

T.C.
İSTANBUL NİŞANTAŞI ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**SOSYAL MEDYA PLATFORMLARINDA YAPAY
ZEKA VE MAKİNE ÖĞRENİM TEKNİKLERİNİ
KULLANARAK, DOĞAL DİL İŞLEME İLE HAKARET
İÇEREN CÜMLE TESPİTİ VE DUYGU ANALİZİNİN
ÖLÇÜLMESİ**

YÜKSEK LİSANS TEZİ

Can Berk ÇELİK

**Enstitü Anabilim Dalı : Bilgisayar Mühendisliği
Enstitü Bilim Dalı : Bilgi Teknolojileri**

Tez Danışmanı: Dr. Öğr. Üyesi Fatih ŞAHİN

MART 2023

T.C.
İSTANBUL NİŞANTAŞI ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**SOSYAL MEDYA PLATFORMLARINDA YAPAY
ZEKA VE MAKİNE ÖĞRENİM TEKNİKLERİNİ
KULLANARAK, DOĞAL DİL İŞLEME İLE HAKARET
İÇEREN CÜMLE TESPİTİ VE DUYGU ANALİZİNİN
ÖLÇÜLMESİ**

YÜKSEK LİSANS TEZİ

ORCID ID: 0000-0003-1198-7771

Can Berk ÇELİK

Enstitü Anabilim Dalı : Bilgisayar Mühendisliği
Enstitü Bilim Dalı : Bilgi Teknolojileri

“Bu tez 08/03/2023 tarihinde aşağıdaki jüri tarafından Oybirliği / ~~Oyçokluğu~~ ile kabul edilmiştir.”

JÜRİ ÜYESİ	KANAATİ	İMZA
Dr. Öğr. Üyesi Fatih ŞAHİN	BAŞARILI	
Dr. Öğr. Üyesi Mahsa TORKAMANIAN AFSHAR	BAŞARILI	
Dr. Öğr. Üyesi Negar Sadat SOLEIMANI ZAKERİ	BAŞARILI	

Enstitü Müdürü
Onay

BEYAN

Bu tezin yazılmasında bilimsel ahlak kurallarına uyulduğunu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez çalışması olarak sunulmadığını beyan ederim.

Can Berk ÇELİK

08/03/2023



ÖNSÖZ

Tez sürecinde yapılan çalışmayı özenle takip eden değerli danışmanım Dr. Öğr. Üyesi Fatih ŞAHİN'nin çok değerli bilgi ve tecrübelerini benimle paylaştığı için kendisine teşekkürlerimi ve saygılarımı sunarım.

Tez çalışmamda bilgi ve tecrübeleriyle en baştan bu zamana kadar desteği olan Aynur Günay arkadaşşıma ve katkısı olan tüm arkadaşlarıma çok teşekkür ederim.

Bu zamana kadar üzerimde çok emekleri olan anneme, babama ve abime çok teşekkür ederim.

Son olarak tez sürecinde her türlü fedakârlığı gösteren eşim Fatma Gül'e ve çocuklarım Anıl ve Alp'e çok teşekkür ederim.

Can Berk ÇELİK
08/03/2023

İÇİNDEKİLER

KISALTMALAR	iii
TABLO LİSTESİ	iv
ŞEKİL LİSTESİ.....	v
GİRİŞ	1
Çalışmanın Yöntem ve Metodu.....	2
Çalışmanın Önemi ve Amacı.....	2
BÖLÜM 1: BİLGİ VE VERİ TEKNOLOJİLERİ.....	7
1.1 Bilgi ve Veri Kavramları	7
1.2 Çok Tercih Edilen Veritabanı Yönetim Sistemleri	7
1.2.1 Oracle.....	7
1.2.2 Mssql.....	8
1.2.3 Postgresql	8
1.3 Verinin Bilgiye Dönüşümü	9
1.4 Veri Madenciliğinde Kullanılan Teknikler	10
1.4.1 Veri Madenciliğinin Kullanıldığı Alanlar	10
1.4.2 Veri Madenciliğinde Karşılaşılan Problemler.....	11
BÖLÜM 2: YAPAY ZEKÂ MAKİNE ÖĞRENİMİ, DOĞAL DİL İŞLEME TEKNOLOJİLERİ	15
2.1 Yapay Zeka Kavramın Nedir, Neden Önemlidir?.....	15
2.2 Yapay Zekâda Kullanılan Ajanlar	18
2.3 Makine Öğrenmesi Teknikleri.....	19
2.3.1 NB – Naive Bayes Algoritması.....	20
2.3.2 SVM – Destek Vektör Makinesi	21
2.3.3 KNN – K En Yakın Komşu	23
2.3.4 LR - Lojistik Regresyon Algoritması.....	24
2.3.5 RF - Rastgele Orman Algoritması	25
2.3.6 ANN - Yapay Sinir Ağları	26
2.4 Makina Öğrenme Algoritmaları Kümeleme.....	29
2.4.1 K-Means Algoritması.....	30
2.5 Doğal Dil İşleme Nedir, Neden Önemlidir.....	32

2.6	Doğal Dil İşlemenin Kullanım Alanları Ve Analiz Edilmesi.....	33
BÖLÜM 3: SOSYAL MEDYA PLATFORMLARI VE TARİHÇESİ		36
3.1	Sosyal Medya Kavramları Ve Tarihçesi	36
3.2	Sosyal Medyada En Çok Kullanılan İçerik Paylaşımlı Siteler:.....	37
3.2.1	Youtube	38
3.2.2	İnstagram.....	38
3.2.2	Sosyal Medyada En Çok Kullanılan Sosyal Ağlar:	38
3.4	Sosyal Medya Neden Önemlidir	40
3.5	Sosyal Medyada En Çok Kullanılan Platform Olan Twitter’ın Yaptığımız Araştırma Açısından Literatüre Katkısı	41
3.6	Sosyal Medya Analitiği Üzerinden Verinin Önemi	42
BÖLÜM 4: BULGULAR.....		44
4.1	Metotların ve Materyallerin Kullanılması	44
4.2	Geliştirmede Kullandığımız Gerekli Kütüphaneler.....	45
4.3	Geliştirmede Kullandığımız Gerekli Kütüphanelerin Açıklamaları.....	46
4.4	Geliştirmede Kullanılan Kodların Açıklamaları	51
SONUÇ.....		66
KAYNAKLAR		68

KISALTMALAR

NLP	: Natural Language Processing
AI	: Artificial Intelligence
IDE	: Integrated Development Environment
YZ	: Yapay Zeka
CRM	: Customer Relationship Management
SQL	: Structured Query Language
PL/SQL	: Procedural Language/Structured Query Language
MSSQL	: Microsoft SQL Server
OCR	: Optical Character Recognition
RAC	: Real Application Clusters
KDS	: Karar Destek Sistemleri
DCS	: Distributed Configuration Store
HA	: High Availability
DBMS	: Database Management System
WCSS	: Within-Cluster Sums of Squares
DDI	: Doğal Dil İşleme
NB	: Navie Bayes
KNN	: K-Nearest Neighbors
SVM	: Support Vector Machine
DT	: Decision Tree
ANN	: Artificial Neural Network
LR	: Logistic Regression
RF	: Random Forest

TABLO LİSTESİ

Tablo 1 Biyolojik sinir sistemi elemanları ve yapay sinir sisteminde karşılıkları	28
Tablo 2 Veri seti alan bilgileri.....	52
Tablo 3 Veri setindeki mevcut durum üzerinden bilgileri	53



ŞEKİL LİSTESİ

Şekil 1: Accuracy value.....	4
Şekil 2: Sosyal Büyük Verinin Yapısı.....	5
Şekil 3 : İşlem Adımları	6
Şekil 4: Veri madenciliğine entegreli olan disiplinler	9
Şekil 5: Destek vektör makinesi	22
Şekil 6: K en yakın komşu.....	23
Şekil 7: Biyolojik sinir hücresi ve yapay sinir ağı.....	27
Şekil 8: Yapay sinir hücresi.....	28
Şekil 9: K-Means algoritması – Rastgele nokta seçilmesi.....	30
Şekil 10: K-Means algoritması – Merkez noktasına göre uzaklık.....	30
Şekil 11: K-Means algoritması – Kümelenen elemanlar	31
Şekil 12: K-Means algoritması – Yeni kümeleme modeli.....	31
Şekil 13: K-Means algoritması – Tekrar edilen adımlar	32
Şekil 14: WCSS denklemi	32
Şekil 15: Geliştirme ortamında kullanılan Anaconda navigator.....	44
Şekil 16: Uygulama aşamaları diyagramı.....	44
Şekil 17: Değer alanları	51
Şekil 18: Veri setinin ilk 10 satırı.....	52
Şekil 19: Veri setinde target kolonunun söylemlere göre görselleştirilmiş dağılımı.....	53
Şekil 20: Veri setinde target kolonunun söylemlere göre numerik dağılımı	54
Şekil 21: Veri setindeki alanların total duruma göre görülmesi	54
Şekil 22: Veri setinin korelasyon matrisine göre değerleri	55
Şekil 23: Veri setindeki alanlara göre korelasyon matrisinin görselleştirilmiş hali	55
Şekil 24: Threshold değeri girilerek korelasyon matrisinin görselleştirilmiş hali	56
Şekil 25: Veri setindeki features'lerin histogram olarak görünümü.....	57
Şekil 26: dt_transformed tablosunun ilk 10 satırı.....	58

İstanbul Nişantaşı Üniversitesi, Lisansüstü Eğitim Enstitüsü Yüksek Lisans Tez Özeti

Tezin Başlığı: Sosyal medya platformlarında yapay zekâ ve makine öğrenim tekniklerini kullanarak, doğal dil işleme ile hakaret içeren cümle tespiti ve duygu analizinin ölçülmesi	
Tezin Yazarı: Can Berk ÇELİK	Danışman: Dr.Öğr.Üyesi Fatih ŞAHİN
Kabul Tarihi: 08.03.2023	Sayfa Sayısı: vi (önkısım)+81(tez)
Anabilim Dalı: Bilgisayar Mühendisliği	Bilim Dalı: Bilgi Teknolojileri
Sosyal medya platformları yaygın olduğu andan itibaren devamlı olarak ilgi odağı olmuştur. Özellikle ülkemizde yaşanan her olay anında sosyal medya platformlarına düşmüş ve bu platformlar sayesinde insanların bilgilenmiş olup, bilgilendiği yerlerden de paylaşımlar yapıyor duruma gelmiştir. Aynı zamanda gerekli platformlara da yorum yapabilme durumları da doğmuştur. Bizleri ilgilendiren kısımda aslında yapılan yorum kısımları çünkü yapılan paylaşımlar sonucu insanlar bu paylaşımı iyi şekilde aktardığı gibi kötü şekilde yani manipüle edip aktarabilir. Burada bizim için önemli olan kısım yapılan yorumları tespit edip, bu yorumlardan hakaret içerikli cümleleri ayıklamak ve değişen duygu durumu sonucunda ise insanlarda duygu analizini belirlemek olacaktır. Sosyal medyada kullanılan, popüler platformların başında gelen twitter platformu üzerinden çalışma yapmayı hedeflemekteyiz. Ayrıca bununla beraber son yıllarda teknolojik gelişmeler ve insanların teknolojik cihazlara erişimleri hızla artmıştır. Özellikle yapay zeka teknolojilerinin kullanılması açısından hızlı bir gelişim olmuş ve sosyal medyada üye olan kullanıcı sayısı artmış. Bu durumla birlikte akıllı telefon, tablet gibi elektronik ürünleri kullanan insan sayısı da hızla yükselmiştir. Bu teknolojik cihazlarla geçirilen zaman arttıkça, sosyal medya platformlarında geçirilen zaman çoğalmış olup, istatistiklere göre daha da çok artmıştır. Hatta birçok insan bu platform üzerinden para kazanmaktadır. Bu platformların başında da twitter, facebook, instagram ve tiktok gibi sosyal medya platformları gelmektedir. Bu alanlarda yapılan görsel ve metinsel paylaşımlar insanlar arasında hızla yayılmakta ve birçok kişiyi etkisi altına almaktadır. Bu kapsamda yapay zeka makine öğrenim teknikleri kullanarak, görüntü işleme ve doğal dil işleme faaliyetleri artmıştır. Şimdi dünyanın bir ucunda gelişen gelişmelerden haberdar olduğumuz gibi yanı başımızda gerçek dışı bilgilere de maruz kalmaktayız. Sosyal medya platformları tek yönlü olmayıp yorum, beğeni, paylaş gibi dinamik bir yapı haline gelmiştir. Yapay zeka ve makine öğrenim tekniklerini kullanarak kullanıcıların duygu düşünce analizi ya da hakaret içeren, saldırgan ve herhangi bir tepkisi olmayan ya da olumlu yorumların tespit edilmesini hedefleyip, doğan ihtiyaçları gidermeye çalışmaktayım. Bunun haricinde zamanla daha da iyi duruma gelecektir. Yapılan tez çalışmasında, twitter üzerinde etiketlenmiş veri setini ve belli sınıflandırma algoritmaların kullanarak, tahmin edilmesi ön görülmektedir. Hangi algoritmayı kullanmak, tez çalışması kapsamında uygundur? Şeklindeki soruları tespit etmek için çalışma yapılması hedeflenmektedir. Aynı zamanda veri seti üzerinde doğruluk skorlarını kullanılacaktır. Bu bağlamda algoritmalarda kullanılan modellere göre doğruluk skorlarının belirlenmesi hedeflenmektedir.	
Anahtar Kelimeler: Hakaret İçerikli Cümleler, Duygu Analizi, Yapay Zeka Teknolojileri, Makine Öğrenmesi, Doğal Dil İşleme	

İstanbul Nişantaşı University, Graduate Education Institute Abstract of Master's Thesis

The Title of Thesis: Using artificial intelligence and machine learning techniques on social media platforms, natural language processing and detection of insulting sentences and measuring sentiment analysis

Author: Can Berk ÇELİK

Supervisor: Assist. Prof. Dr. Fatih ŞAHİN

Date: 08.03.2023

Nu. Of Pages: vii (pre text)+81(main body)

Department Dalı: Computer Engineering

Subfield: Information Technologies

Social media platforms have been the focus of attention ever since they became widespread. Especially, every event in our country has instantly fallen onto social media platforms, and thanks to these platforms, people have been informed and they have been sharing from the places where they are informed. At the same time, it is possible to comment on the necessary platforms. The part that interests us is actually the comment sections, because as a result of the sharing, people can transfer this share in a good way as well as in a bad way, that is, they can manipulate and transfer it. Here, the important part for us will be to identify the comments made, to extract the insulting sentences from these comments, and to determine the changing sentiment analysis in people as a result. We aim to work on the twitter platform, which is one of the most popular platforms used in social media. Also, however In recent years, technological developments and people's access to technological devices have increased rapidly. Especially in terms of the use of artificial intelligence technologies, there has been a rapid development and the number of users on social media has increased. Along with this situation, the number of people using electronic products such as smart phones and tablets has increased rapidly. As the time spent with these technological devices has increased, the time spent on social media platforms has increased and has increased even more according to statistics. Many people even make money on this platform. At the beginning of these platforms are social media platforms such as twitter, facebook, instagram and tiktok. Visual and textual sharings in these areas spread rapidly among people and affect many people. In this context, image processing and natural language processing activities have increased by using artificial intelligence machine learning techniques. As we are aware of the developments in one end of the world, we are also exposed to unrealistic information right next to us. Social media platforms are not one-sided but have become a dynamic structure such as comments, likes and shares. By using artificial intelligence and machine learning techniques, I aim to identify users' sentimental analysis or offensive, offensive and unresponsive or positive comments, and try to meet the arising needs. Other than that, it will get better over time. In the thesis study, it is predicted to be estimated by using the tagged data set on Twitter and certain classification algorithms. Which algorithm is suitable for the thesis study? It is aimed to carry out a study to determine the questions in the form. At the same time, accuracy scores will be used on the data set. In this context, it is aimed to determine the accuracy scores according to the models used in the algorithms.

Key Words: Offensive Sentences, Sentiment Analysis, Artificial Intelligence Technologies, Machine Learning, Natural Language Processing

GİRİŞ

Sosyal medya platformlarından twitter platformunu analiz edeceğiz. Twitter platformunda geçen hakaret içerikli cümleleri tespit edip insanların duygularını ölçererek ülkemiz için oluşturan tehdit unsurlarını ortadan kaldırmayı hedeflemekteyiz. İnternetin yaygın kullanımı insanlara düşüncelerini ifade etmeleri için daha fazla özgürlük vermektedir. Özellikle sosyal medya platformlarında insanlar düşüncelerini özgürce ifade etmektedir. Twitter, Facebook, İnstagram gibi sosyal medya platformları bu kategorilerin en başında gelmektedir. Ayrıca internetin artması ile birlikte farklı kültürden, ırktan, dinden, kökenden birçok insanlar arasındaki etkileşim arttı. Bu şekilde etkileşimin artması ile sanal ortamda bilgiye ulaşmak daha kolay hale gelmiştir. Birkaç saniye içinde insanlar istedikleri paylaşıma ulaşabiliyorlar. İnsanların kendini ifade etmesi için yeni bir fırsatlar galaksisi açtığı için içerik fikirlerinin çevrimiçi olarak olumlu bir şekilde fikir alışverişine yol açmasının yanı sıra nefret söyleminin yayılmasına olumsuz bir şekilde katkıda bulunmaktadır. Sosyal medya platformlarında Anonim olmak, İnternetin sağladığı kullanıcı sayısında çoğalmalara sebep olmakla beraber bilinmezliği de arttırmaktadır. Aynı zamanda toplumu da etkileyerek web içeriklerini küfürlü kelimeler, nefret söylemi platformlarda toplumu zedeler bu yüzden önüne geçmek için hakaret içerikli cümleleri ayıklamamız öngörülmektedir. İnsanların din, dil, ırk, cinsiyet ayırt etmeden psikolojik olarak da sosyal medyadan gerekli istatistiksel sonuçlara göre durumlarını analiz etmeyi hedeflemekteyiz. Her geçen gün sosyal medya platformlarında ki kullanıcı sayıları giderek artmaktadır. Özellikle bazı kullanıcılar kendilerine gerçeği yansıtmayan hesap açarak top trend olmuş bir konu ile ilgili hakaret içerikli cümlelerde bulunabilme potansiyeli yüksek kullanıcılar bulunmaktadır. Bunların önüne geçmek için tezde yaptığımız çalışmayı geliştirip hakaret içerikli cümleleri tespit edildiği gibi hangi kullanıcının da söylediği tespit edilebilir. Bu bağlamda adli mercilerin bilgisine sunularak sosyal medyada geçen suç oranı düşürülmesi ön görülebilir. Bu çalışmamızın en büyük hedeflerinden bir tanesi de sosyal medyada geçen küfürlü içerikleri tespit edip, insanların nasıl düşündüğüyle ilgili psikolojik analiz çalışması sonucunda duygu analizini de ölçüyor duruma gelmiş olmayı hedeflemekteyiz. İnternet yaygın olduğu sürece sosyal medya platformlarında hakaret içerikli söyleme hazır kullanıcılar devamlı artacağı ön görülmektedir. Neticesinde de

kullanılacak veri seti de çıkan yorumlara göre şekillenmeye devam edecektir. Hakaret içerikli söylemler ve Saldırgan tavırlı cümlelerin tespiti sonucundan çıkan duygu analizine göre insanların davranışlarının yapay zekâ ve makine öğrenmesi teknikleriyle nasıl olacağı ön görülebilir. Önemli olan doğru sınıflandırma algoritmalarını kullanarak gerekli yöntemleri yapmak.

Çalışmanın Yöntem ve Metodu

Literatür taramasında, çalışma yapılan konuyla ilgili akademik yayınlar, yöktez, google scholar gibi akademik platformlar incelenmiş, konuyla ilgili hakaret içerikli cümle tespitlerin çeşitli tezlerde, makalelerde belli alanlarla ilişkilendirerek akademik yayınlar ortaya çıkmıştır. Bizim yapmış olduğumuz sosyal medya platformlarında yapay zeka makina öğrenimi tekniklerini kullanarak hakaret içerikli cümleleri tespit ettikten sonra sosyal medya platformlarında insanların daha özgür bir ortamda hakaret ortamı olmadan düşüncelerini kolayca ifade etmelerini sağlamak. Twitter, Facebook, İstagram gibi sosyal medya platformlarının özellikle twitter platformunu kapsayacak şekilde değerlendirecek, insanların din, dil, ırk, cinsiyet ayrımı yapmadan etkileşimlerini desteklemek için hakaret içeren sözcüklerin de önüne geçmek için bu tezi yazmayı uygun gördük. Yapılan araştırmalar sonucu da sosyal medyanın ihtiyaçlarına binaen insanların kendilerini ifade etmeleri için yeni fırsatlar dünyası olarak nitelendirilebilir.

Tez çalışmasında yapay zeka ve makine öğrenim teknikleri uygulanmıştır. Bu kapsamda veri seti paylaşım platformlarından bir tanesi olan kaggle platformundan edinmiş olduğumuz veri setini, geliştirdiğimiz ortamda hakaret içeren cümleler, saldırgan tavırlı cümleler ve herhangi bir etkisi olmayan cümleleri tahmin edip görmemizi sağlar.

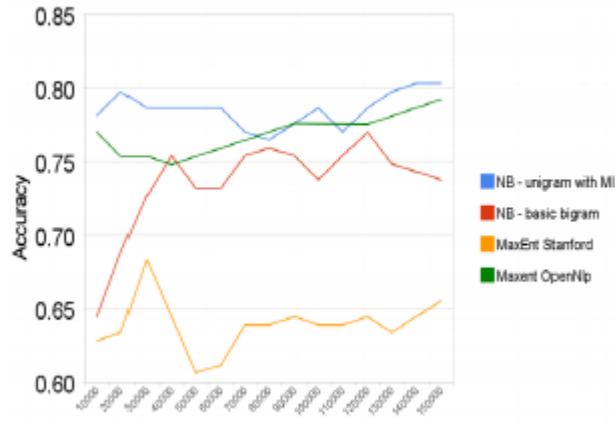
Çalışmanın Önemi ve Amacı

Çalışmamızın amacı hakaret içerikli cümleleri kullandığımız IDE aracılığıyla preprocessing işlemi yaparak yazdığımız cümlelerin kullandığımız algoritmalar arasında 1'e yakın olup kuvvetle muhtemel tahmin etme kapasitesi hangisi yüksek ise o algorithmadan işlemlerini yapıyoruz. Neticesinde bize hakaret içerikli mi, Saldırgan mı yoksa ne hakaret içerikli ne de saldırgan mı? Diye etiketlenmiş verimiz sayesinde bize sonuç dönüyor ve duygu analizi bu sonuçlara göre yapılıyor. Teknik detaylara bulgular tarafında değinilecek. NLP ile ilgili çeşitli araştırmalar yapılmış, makaleler yazılmıştır.

Çalışmada kullanılan yöntem ise Yapay zeka, doğal dil işleme (NLP) teknolojileri kullanarak hakaret içerikli cümleleri tespit edebilir. Bu teknolojiler, dilin yapısal özelliklerini anlama, kelime anlamlarını anlama ve dilin bağlamını anlama yeteneğine sahip olabilirler. Ayrıca, makine öğrenmesi ve derin öğrenme teknikleri de kullanılarak, yapay zeka modelleri, daha geniş bir veri kümesinde öğrenerek, hakaret içeren ifadeleri tespit etmek için daha iyi hale gelebilirler. Hakaret içerikli cümleleri tespit etmek için kullanılan yöntemler arasında, kelime frekansı tabanlı yaklaşımlar, kelime gömme (word embedding) ve sinir ağları temelli modeller yer alır. Kelime frekansı tabanlı yöntemler, bir cümle içindeki belirli kelimelerin belirli bir dilim içinde ne kadar sıklıkla kullanıldığını hesaplayarak, cümlelerin hakaret içerip içermediğini belirler. Bu yöntemlerin dezavantajı, cümledeki kelime sırasının ve bağlamın göz ardı edilmesidir. Kelime gömme (word embedding), her kelime için benzersiz bir vektör oluşturur ve bu vektörlerin benzerliğine dayalı olarak hakaret içerikli cümleleri tespit eder. Bu yöntem, kelime sırasını ve bağlamı daha iyi anlar ve daha iyi sonuçlar verir. Sinir ağları temelli modeller, büyük veri setleri kullanarak öğrenir ve daha kesin sonuçlar elde etmek için diğer yöntemlere göre daha karmaşık bir yapıya sahiptir. Örneğin, önceden eğitilmiş bir sinir ağı modeli, bir cümleyi analiz ederek hakaret içeren kelimeleri tespit edebilir. Ancak, hakaret içerikli cümleleri tespit etmek için kullanılan yöntemlerin doğruluğu, önceden belirlenmiş bir standart ile karşılaştırılarak ölçülmelidir. Ayrıca, bu yöntemlerin kullanıldığı platformların, kullanıcıların verilerinin gizliliğini koruma konusunda dikkatli olmaları önemlidir.

Bu konuda ilk araştırmalar 2000'li yılların hemen öncesinde başlamış olsa da asıl sonuç alınan çalışmalar bu milenyumdan sonra elde edilmiştir. Duygu analizi alanında yapılan ilk çalışmalardan biri olarak kabul edilen “Thumbs up?”, Bo Pang, Lillian Lee ve Shivakumar Vaithyanathan tarafından 2002 de yayınlanmıştır. Film yorumlarını kendilerine metin olarak almışlardır ve bunları unigram, bigram, unigram+bigram, part of speech gibi makine öğrenmesi yöntemleriyle pozitif, negatif olarak ayırmışlardır. Daha karışık yapıda olan cümleleri, (örn: “Bir insan 2 saat boyunca nasıl bu filmi izleyebilir” gibi) hiçbir negatif, pozitif belirleyici kelime olmamasına rağmen cümlelerin hangi duyguyu taşıdığı sonucuna ulaşmak istenmiştir. Makine öğrenmesi metodlarından naive bayes, maximum entropi ve Support vector machine algoritmalarında başarı elde etmişlerdir. En iyi sonucu ise unigram özelliğine göre makine öğrenmesi ayarlandığında

alınmıştır. Svm algoritmasının performansı %82 lere ulaşmıştır. 2009’da duygu analizini Twitter için uyarlayan Alec Go, Lei Huang, Richa Bhayani mesajları sınıflandırma üzerine uğraşmışlardır. Uzaktan denetimli öğrenme algoritmasını kullanarak 1,6 milyon tweeti incelemeye almışlardır. Pang ve grubunun yaptığı Thumbs up? Projesini kendilerine kısmi olarak bir kaynak almışlardır. Bu çalışmayı unigram, bigram, unigram+bigram olarak analizini gerçekleştirmişlerdir. Şu şekilde bir sonuç elde etmişlerdir (Nalçakan ve diğerleri,2015:3).

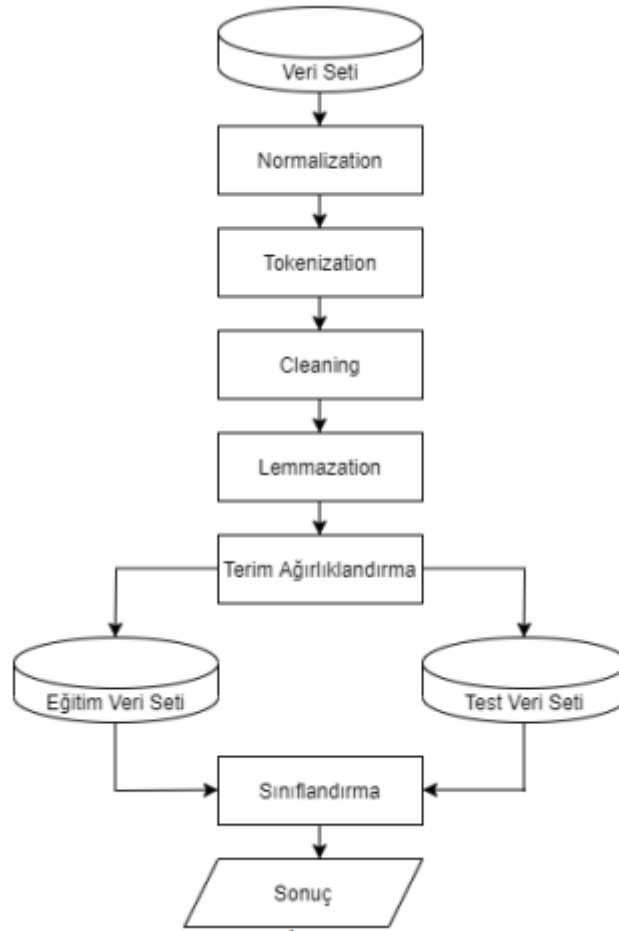


Şekil 1: Accuracy value

Kaynak: Nalçakan, Bayramoğlu ve Tuna,2015:4

Burada genel olarak en yüksek sonucu naive bayes algoritmasının unigramın karşılıklı bilgi biçiminde kullanılmasıyla elde edilmiştir. Ancak grafikte gösterilmeyen unigramın support vector machine algoritmasıyla kullanılması %82 lere bir doğruluk sonucu doğurmuştur (Nalçakan ve diğerleri,2015). Yapılan bu çalışmaya göre kullanılan yöntemler arasında geleneksel ve eğitilmiş yöntemler kullanılmıştır. Kullanılan bu yöntemlere göre ise gerçekleşen deneyler arasında accuracy value de belirtilen algoritmalar kullanılarak grafik oluşmuştur. Yine yapılan çalışmaya atıf yapmak gerekirse yaptığımız tez ile benzer nitelikler taşımaktadır. Totalde belirtilen veri setini eğiterek, etiketlenmiş veri seti haline getirdikten sonra çıkan tweet sayısı kadar kullanılan algoritmalarla pozitif, negatif ya da nötr olarak ayrılmış cümleler gözükmektedir.

Sosyal medya verisi büyüklüğü, zenginliği, dinamikliği, insan davranışını temsil başarısı ile bireyleri, grupları ve toplumları anlama imkânı sunmaktadır (Batrince ve Treleaven, 2015). Teknolojik gelişmeler, sosyal medya platformlarında oluşturulan



Şekil 3 : İşlem Adımları

Kaynak: Kumaş,2021:2

Şekil 3 diyagramına göre her bir işlem belli aşamalarda ki algoritmalara tabi tutularak sonuç elde edilmiş olur. Sosyal medya platformlarında hakaret içerikli cümle tespiti ve duygu analizi çalışması da şekil 3'te ki süreçler belli aşamalardan geçilerek sonuca varmaya hedef almıştır.

BÖLÜM 1: BİLGİ VE VERİ TEKNOLOJİLERİ

1.1 Bilgi ve Veri Kavramları

Veri, bilgiye dönüşmeden önceki ham halinde olan kısmına denir. Veri o kadar önemlidir ki bilgiye dönüştürünce her türlü fayda sağlayabilir. CRM (Customer Relation Management) müşteri ilişkileri yönetimini baz alırsak bir ürünü almak için gelen müşterinin tüm bilgilerini bu veriler üzerinden çektiği için müşteriye ürünü satmak ya da cazip gelmesi bu denli önemlidir. Bu yüzden de büyük veri günümüzde önemli rol oynamaktadır. Verilerin işlenerek anlamlı hale geldiğinde bilgiye dönüşür. Verilerin depolandığı alanlara da veritabanı yönetim sistemi alanları (database management system) denir. Bu alanlarda bilgiye ulaşmak için veritabanı yönetim sistemler içinde belli prosedürlerle yapısal olarak yapılandırılmış sorgu dillerini kullanarak bilgiye kolayca ulaşmamızı sağlar. Özellikle günümüz teknolojilerinde maliyeti azaltarak açık kaynak kodlu sisteme geçilmektedir. Genellikle maliyeti yüksek olan profesyonel teknolojiler hala kullanılmaktadır.

1.2 Çok Tercih Edilen Veritabanı Yönetim Sistemleri

1.2.1 Oracle

Küresel olarak en çok tercih edilen, profesyonel anlamda en çok kullanılan ilişkisel veritabanı yönetim sistemlerinin başında Oracle gelmektedir. Oracle hız ve güvenlik anlamında bankalarda, savunma sanayii'lerinde ve askeri kurumlarda genellikle tercih edilmektedir. Oracle'ın en büyük avantajı ise aktif – aktif bir yapı sunmasıdır. Bu sayede sunuculardan birisi kullanılmaz duruma geldiğinde diğer sunucu devreye girer bu sayede herhangi bir hizmet kesintisi olmadan işlemler devam eder. Lisans maliyetleri ise yüksektir. Oracle veritabanı mimarisine göre yüksek kullanılabilirlik (high availability) yapısı olduğundan dolayı RAC yapısı içinde hizmet sunar.

Oracle Database, büyük ölçekli kurumsal uygulamalar için tasarlanmış bir veritabanıdır ve birçok özellik sunar, örneğin yüksek kullanılabilirlik, güvenlik, performans, veri analizi ve yönetim gibi seçenekler sunar. Oracle Database, SQL (Structured Query Language) sorgularını ve PL/SQL (Procedural Language/Structured Query Language) programlama dilini kullanarak yönetilir. Oracle Database, Windows, Linux, Unix ve diğer işletim sistemleri üzerinde çalışabilir. Oracle Database, farklı sürümleri ile mevcuttur, örneğin Standard Edition, Enterprise Edition ve Express Edition gibi. Her

sürüm, farklı özellikleri ve kapasiteleri ile birlikte gelir ve bu nedenle, kullanım senaryolarına göre seçilebilir. Oracle Database, birçok endüstride yaygın olarak kullanılmaktadır ve büyük kuruluşlar ve hükümet kurumları gibi büyük müşteri tabanlarına sahiptir.

1.2.2 Mssql

Microsoft SQL Server, bir ilişkisel veritabanı yönetim sistemi (RDBMS) olup, Microsoft tarafından geliştirilmiştir. SQL Server, büyük kurumsal veritabanları ve iş uygulamaları için tasarlanmıştır ve birçok özellik sağlar, örneğin çok kullanıcılı destek, yedekleme ve kurtarma, işlem yönetimi, yüksek kullanılabilirlik, veri analizi ve raporlama, güvenlik gibi seçenekler sunar. SQL Server, Transact-SQL (T-SQL) olarak adlandırılan bir SQL varyasyonunu kullanarak yönetilir. SQL Server, Windows işletim sistemleri için tasarlanmıştır ve Windows Server işletim sistemi üzerinde çalışabilir. SQL Server, farklı sürümleri ile mevcuttur, örneğin Express, Standard, Enterprise, ve Developer gibi. Her sürüm, farklı özellikleri ve kapasiteleri ile birlikte gelir ve bu nedenle, kullanım senaryolarına göre seçilebilir.

Genellikle kurumun hizmet verdiği alana göre tercih edilir. Kurumlarda kullanım alanına göre lisans maliyeti olan ve lisans ücreti alınıp kurulan bir ilişkisel veritabanı yönetim sistemidir. Genellikle kurumlarda web platformlarına hizmet vermesi amacıyla kullanılır. Hızlı ve güvenilir olması sebebiyle de tercih sebebi olur.

1.2.3 Postgresql

Günümüzde en yaygın olarak görülen, lisans maliyeti olmayan, pek çok kurumda yaygınlaşan açık kaynak kodlu veritabanı popüler hale gelmiştir. Verdiği hizmetlerle aktif-pasif çalışarak yedeklilik sağlamaktadır. Bu yapıya ek olarak patroni sistemi kurumlarda kullanılmaktadır. Patroni yapısının çalışması anlamında otomatik diğer sunucuya kendisini lider olarak tanıtmayı için DCS araçlarını kullanması gerekmektedir. ETCD, Consul, Kubernetes gibi dcs araçlarından birini kullanarak otomatik failover yapısı kullanılır. Bu sayede örnek vermek gerekirse 3 sunucuda patroni yapısı var. Birincisi primary, diğer ikisi ise slave patroni sayesinde birinci sunucu hata verdiğinde diğer ikisinden birinde yük hangisinde az ise devam eder bu sayede hizmet kesintisi yaşanmaz. Hata veren sunucu ise replica duruma geçer yani primary(lider) sunucuya yazılan verilere diğer sunuculara replica(aktarılır) edilir. Diğer sunucular sadece veriyi

okuyabilir. Postgresql de HA yapısına sahiptir. Cluster yapısının oluşturulması ve idame edilebilmesinde kullanılan bir açık kaynak aracıdır. Python dili ile geliştirilmiştir. Postgresql cluster kurulum, replikasyon, otomatik failover yapılması anlamında bu amaçla kullanılmaktadır.

1.3 Verinin Bilgiye Dönüşümü

Sosyal medya platformları olsun, müşteri giyim ürünleri, restoranlar gibi yerler de artık o kadar veri girişi oluyor ki bir süreden sonra anlamlı anlamsız veriler artık birikiyor. Örneğin bir satış mağazasında kullanılan CRM (Customer Relationship Management – Müşteri İlişkileri Yönetimi) sayesinde müşterinin bilgisi veritabanına kaydedildikten sonra artık müşteriye ait beden yapısı, aldığı üründen sevdiği renge kadar birçok çıkarım elde ediliyor. Bu veriler sayesinde müşteriye indimler ya da farklı bir rol üstlenerek müşterinin odağını çekmeye çalışıyorlar. Bunun neticesinde ise müşterilerin mağazaya sadakatleri de ölçülebiliyor. Veriler ham halde anlamlı hale gelmediği sürece aktif rol oynamaz ama veriler işlenerek bilgiye dönüşür ve anlamlı hale gelir ise saydığımız ve saymadığımız bu sektörlerde de kullanılan yerler için büyük rol oynar. Büyük veri de bu şekilde ortaya çıkmıştır. Anlamlı anlamsız çok büyük hacimli veriler olduğundan dolayı bu verileri ayıklamak gerekiyor. Anlamlı hale getirmek için de veri madenciliği yöntemi burada devreye giriyor. Veri madenciliği birçok disiplinle etkileşim halindedir. Bu disiplinlerin tam merkezinde yer alır.



Şekil 4: Veri madenciliğine entegreli olan disiplinler

Kaynak: Savaş,Topaloğlu ve Yılmaz,2012:3

Verinin bilgiye dönüşümünden sonrasında bazı karar mekanizmalarının aktif rol oynayabilmesi için ve işletmelerde efektif, etkin ve doğru kararlar almasına yönelik KDS karar destek sistemleri ortaya çıkmıştır. Geçmişte baktığımızda kullanılan karar destek sistemleri klasik yapıda kullanılan araçlardan farklı olarak, veri madenciliğinde fazlasıyla kapsamlı ve yapının otomatize edilmiş hali ile servis sunulması hedeflenmiştir. Buradaki karar destek sistem yapısının amacı ise aktif olarak pazarlama süreçlerini kullanmaktır. Şekil 2’ de görüldüğü gibi veri madenciliği, diğer unsurlarla entegreli olmak zorundadır çünkü veri toplayıp bilgiye dönüştürme anlamında diğer disiplinlerle kendisini beslemek zorundadır.

Verinin bilgiye dönüşümü, veri analizi, yorumlama ve anlamlandırma işlemleri yoluyla gerçekleştirilen bir süreçtir. Veri, anlamlı hale getirilene kadar sadece sayılar, metinler ve semboller gibi anlamsız bilgilerdir. Ancak, bu veriler işlenerek bilgiye dönüştürülebilirler. Verinin bilgiye dönüştürülmesi, işletmelerin karar vermelerine yardımcı olur. Veri analizi ve raporlama, bir işletmenin faaliyetlerini ve performansını anlamasına yardımcı olabilir, böylece işletmeler, karar verirken daha doğru bilgilere sahip olabilirler.

1.4 Veri Madenciliğinde Kullanılan Teknikler

1.4.1 Veri Madenciliğinin Kullanıldığı Alanlar

Büyük verinin olduğu her alanda kullanılması kuvvetle muhtemel mümkündür. Veri her geçen gün artması ile beraber kullanım alanlarını yaygınlaştırıyor. Bu bağlamda birçok alanda artık veri madenciliği yöntemleri kullanılmaktadır.

Büyük hacimde veri bulunan her yerde veri madenciliği kullanmak mümkündür. Günümüzde karar verme sürecine ihtiyaç duyulan birçok alanda veri madenciliği uygulamaları yaygın olarak kullanılmaktadır. Örneğin pazarlama, biyoloji, bankacılık, sigortacılık, borsa, perakendecilik, telekomünikasyon, genetik, sağlık, bilim ve mühendislik, kriminoloji, sağlık, endüstri, istihbarat vb. birçok dalda başarılı uygulamaları görülmektedir (İnan, 2003; Albayrak, 2008; Akgöbek ve Çakır, 2009).

1.4.2 Veri Madenciliğinde Karşılaşılan Problemler

(Savaş, Topaloğlu ve Yılmaz, 2012: 1-23) Verilerin tutulduğu veritabanlarında veriyi işledikten sonra bilgiye dönüştüğü andan itibaren büyük kapasiteli veriler ortaya çıkmaktadır. Bunun sonucunda karşımıza çıkma ihtimali olan bazı problemlerle karşı karşıya kalma ihtimalimiz çok yüksektir. Bu problemlerin bazıları da şunlardır:

Artık veri: Elimizdeki verilerin istenilen problemi çözme sürecinde kullanılan örneklem kümesindeki işe yaramayan gereksiz çıkarımlardır. İş sürecinde belli başlı aşamalarda karşımıza çıkma ihtimali kuvvetle muhtemel yüksektir.

Belirsizlik: Veri madenciliği projelerinde belirsizlik, bir veri kümesindeki verilerin doğruluğu, eksik bilgi durumu, değişkenler arasındaki ilişkilerin belirsizliği ve modele uygulanan algoritmanın doğruluğu gibi faktörlerden kaynaklanabilir. Belirsizlik, modelin doğruluğunu ve sonuçların güvenilirliğini etkileyebilir.

Belirsizlikle başa çıkmak için veri madenciliğinde bazı teknikler kullanılabilir:

İstatistiksel analiz: İstatistiksel analiz, belirsizliği azaltmak için kullanılan en yaygın tekniktir. Örneğin, bir veri kümesindeki belirsizliği belirlemek için ortalama, varyans veya standart sapma gibi istatistiksel ölçümler kullanılabilir.

Olasılık teorisi: Olasılık teorisi, belirsizlikle başa çıkmak için kullanılan bir diğer tekniktir. Bu teknik, belirsizliği olasılık hesaplamalarıyla azaltır. Örneğin, bir satın alma işleminin dolandırıcılık olasılığını hesaplamak için olasılık teorisini kullanabilirsiniz.

Makine öğrenmesi: Makine öğrenmesi, belirsizliği azaltmak için kullanılan bir başka tekniktir. Bu teknik, modelin daha doğru tahminler yapmasına yardımcı olan belirsizliği gidermek için kullanılır. Örneğin, bir modelde eksik veriler varsa, makine öğrenmesi algoritmaları eksik verileri tahmin etmek için kullanılabilir.

Veri görselleştirme: Veri görselleştirme, verilerin daha iyi anlaşılmasına yardımcı olmak için kullanılır. Görselleştirme teknikleri, verilerin özellikleri arasındaki ilişkiyi açıklamak için kullanılabilir.

Belirsizlik, veri madenciliği projelerinde karşılaşılan yaygın bir problem olsa da, yukarıda bahsedilen tekniklerle başarılı bir şekilde azaltılabilir. Hangi tekniklerin kullanılacağı, veri kümesinin özelliklerine, belirsizliğin derecesine ve analiz amacına bağlı olarak belirlenmelidir.

Boş Veri: Kullanılan veritabanında primary key alanı null ise burada alınan değer olma ihtimali vardır.

Dinamik Veri: Dinamik veri, zamanla değişen ve güncellenen verilerdir. Veri madenciliği alanında, dinamik verilerin analizi ve keşfi önemli bir konudur. Dinamik veriler, genellikle bir dizi zaman damgası (timestamp) ile ilişkilendirilir ve zamanla değişen trendleri, kalıpları ve davranışları belirlemek için kullanılır. Dinamik veri kaynakları, sosyal medya, web sayfaları, akıllı cihazlar, sensörler ve diğer otomatik veri toplama sistemleri gibi birçok kaynakta bulunabilir. Örneğin, bir perakende mağazasında, satışların günlük veya saatlik olarak kaydedilmesi, müşteri trafiği, stok seviyeleri ve diğer verilerin analiz edilmesi, dinamik veri analizi için kullanılabilir. Dinamik veri analizi, zaman serisi analizi, desen tanıma, regresyon analizi ve diğer veri madenciliği teknikleri gibi birçok yöntem kullanır. Zaman serisi analizi, zaman içindeki veri değişimlerinin modellemesini sağlar. Desen tanıma, dinamik verilerdeki trendleri ve kalıpları belirlemek için kullanılır. Regresyon analizi, değişen veriler arasındaki ilişkiyi modellemek için kullanılır. Dinamik verilerin analizi, özellikle büyük veri setleri için önemlidir. Bu verilerin analizi, gelecekteki trendleri ve davranışları öngörmek, performansı izlemek, hataları tespit etmek ve iş kararlarını desteklemek için kullanılabilir.

Eksik Veri: Veri madenciliği projelerinde eksik veri problemleri oldukça yaygındır. Eksik veri, veri kümesindeki bir veya daha fazla özellik için eksik değerler içeren gözlemleri ifade eder. Eksik veri, veri kümesinin analizini ve yorumlanmasını zorlaştırabilir ve sonuçların güvenilirliğini azaltabilir.

Eksik veri problemini çözmek için birkaç yöntem vardır. Bunlar arasında:

- **Eksik Veri Analizi:** Bu yöntem, eksik veri problemine neden olan faktörleri belirlemek için kullanılır. Bu faktörler, veri kaynağından, ölçüm hatalarından veya kullanılan veri girişi yöntemlerinden kaynaklanabilir. Eksik veri analizi, eksik veri problemini çözmek için yapılacak diğer işlemler için bir temel oluşturabilir.
- **Eksik Verilerin Silinmesi:** Eksik veri problemini çözmek için en basit yöntem eksik verilerin silinmesidir. Ancak, bu yöntem veri kaybına neden olabilir ve sonuçlar üzerinde önemli bir etkiye sahip olabilir.
- **Değer Atama:** Eksik veri problemini çözmek için daha yaygın bir yöntem, eksik verileri uygun bir değerle doldurmaktır. Bu, ortalama, medyan, mod veya tahmin edilen değerlerle yapılabilir.
- **Eksik Veri Tahmini:** Bu yöntem, eksik verileri tahmin etmek için kullanılan istatistiksel modelleri kullanır. Eksik veri tahmininde, regresyon, çoklu imputasyon ve karar ağaçları gibi yöntemler kullanılabilir.
- **Makine Öğrenmesi Tabanlı Teknikler:** Eksik veri problemini çözmek için makine öğrenmesi tabanlı teknikler de kullanılabilir. Bu teknikler, eksik verileri tahmin etmek için verileri öğrenir ve daha sonra eksik verileri doldurmak için kullanılır. Eksik veri problemini çözmek için hangi yöntemin kullanılacağı, veri kümesinin özelliklerine, eksik veri sayısına ve analiz amacına bağlı olarak belirlenmelidir.
- **Sınırlı Bilgi:** Sınırlı bilgi, veri madenciliği projelerinde sık karşılaşılan bir problem haline gelir. Sınırlı bilgi, bir veri kümesinde eksik bilgi veya yanlış bilgi durumunu ifade eder. Örneğin, bir müşterinin doğru yaşını veya maaşını bilmiyor olabilirsiniz veya bir ürünün geri dönüş oranı hakkında yanlış bir bilgiye sahip olabilirsiniz.

Sınırlı bilgi problemini çözmek için bazı yöntemler vardır:

- **Kaynakları Araştırın:** Bilgiyi tamamlamak için daha fazla kaynak araştırması yapın. İlgili endüstrilerde veya sektörlerdeki diğer uzmanlara danışın veya bir veri kaynağı oluşturmak için anketler yapın.
- **Benzer Veri Kümesi Analizi:** Benzer veri kümesi analizi, eksik bilgiyi tamamlamak için bir yol sağlar. Örneğin, eğer bir müşterinin yaşını bilmiyorsanız, diğer müşterilerin yaşlarına dayanarak müşterinin yaşını tahmin edebilirsiniz.
- **Veri Görselleştirme:** Verilerinizi görselleştirmek, veri analizindeki boşlukları doldurmanıza yardımcı olabilir. Örneğin, bir tablodaki verileri görselleştirerek, eksik verileri daha kolay görebilir ve analiz edebilirsiniz.
- **Makine Öğrenmesi Tabanlı Teknikler:** Makine öğrenmesi tabanlı teknikler, eksik bilgileri tamamlamak için kullanılabilir. Bu teknikler, veri kümesi üzerinde bir model eğitir ve eksik bilgileri tamamlamak için bu modele dayanır.
- **Çoklu Kaynak Kullanımı:** Çoklu kaynak kullanımı, sınırlı bilgi problemini çözmek için farklı kaynaklardan bilgi toplamayı içerir. Bu yöntem, farklı veri kaynaklarından elde edilen verileri birleştirerek eksik bilgileri tamamlamak için kullanılır.

Sınırlı bilgi problemleri, veri madenciliği projeleri için yaygın bir problem olmasına rağmen, yukarıdaki yöntemlerle çözülebilir. Hangi yöntemin kullanılacağı, veri kümesinin özelliklerine, eksik bilgi sayısına ve analiz amacına bağlı olarak belirlenmelidir.

Veritabanı Boyutu: Veritabanı boyutları büyük bir hızla artmaktadır. Veritabanı algoritması çok sayıda küçük örnekleme ele alabilecek biçimde geliştirilmiştir. Aynı algoritmaların yüzlerce kat büyük örneklerde kullanılabilmesi için çok dikkat gerekmektedir.

BÖLÜM 2: YAPAY ZEKÂ MAKİNE ÖĞRENİMİ, DOĞAL DİL İŞLEME TEKNOLOJİLERİ

2.1 Yapay Zeka Kavramın Nedir, Neden Önemlidir?

Yapay zeka, aslında makinaların zeki olması bununla beraber bilgisayar programlarını geliştirme bilimi ya da mühendisliğidir. Makinaya gerekli algoritmalar aracılığıyla öğretmek artık faaliyete geçirme aşamaları vardır.

Genel olarak yapay sinir ağları insan beyninin biyolojik sinir yapısını taklit ederek sinirsel algılayıcılar yardımı ile önceden öğrenilmiş ya da sınıflandırılmış bilgileri kullanarak yeni bilgiler türetebilen ve oluşturabilen, karar verebilen bilgisayar programlarıdır. Örüntü tanıma, sistem tanılama, robotik, sinyal işleme, non lineer denetim alanları gibi birçok uygulama alanında yapay sinir ağları yaygın olarak kullanılmaktadır (Efe, Abadoğlu, & Kaynak, 1999; Efe & Kaynak, 1999; Narendra & Parthasarathy, 1990).

- 1943-McCulloch & Pitts: Beynin Boolean devre modeli
- 1950-Turing'in "Bilgi işleyen makineler ve zekâ"
- 1956-Dartmouth Görüşmesi: "Yapay zekâ" ismi ortaya atıldı.
- 1952-1969-IBM satranç oynayabilen ilk programı yazdı. YZ konusundaki ilk uluslararası konferans düzenlendi.
- 1950ler-İlk YZ programları, Samuel'in kontrol edici programı, Newell ve Simon'ın mantık teoristi, Gelernter'in geometri motoru.
- 1965-Robinson'un mantıklı düşünme için geliştirdiği tam bir algoritma
- 1966-73-YZ hesapsal karmaşayla karşılaşır. Sinir ağları araştırmaları hemen hemen kaybolur.
- 1969-79-Bilgiye dayalı sistemlerin ilk gelişme adımları
- 1980-YZ Endüstri haline gelir
- 1986-Yapay sinir ağları tekrar popüler olur.
- 1987-Yapay Zekâ bilim haline gelir.
- 1995-Zeki ajanlar (terimsel kullanımdır) ortaya çıkar.
- 1997-Deep Blue Kasparov'u yendi.

- 1998-İnternetin yaygınlaşması ile yapay zeka tabanlı birçok program geniş kitlelere ulaştı.
- 2000-05-Robot oyuncaklar piyasaya sürüldü. Halen birçok elektronik cihazda yapay zeka uygulamaları kullanılmaktadır.

Geçmişten günümüze gelen birikimle YZ hakkında şu tanımlamalar yapılmıştır: “Yapay zeka insan tarafından yapıldığında zeki olarak adlandırılan davranışların makine tarafından yapılmasıdır.” “İnsan aklının nasıl çalıştığını göstermeye çalışan bir kuram” “Yapay zeka’nın amacı insan zekâsını bilgisayar aracılığıyla taklit etmektir.” “Yapay zeka makineleri kontrol eden bilgisayar programları oluşturarak zekânın yapısını anlamaya çalışır.” Bu bağlamda yapay zeka birçok alt disiplinlere ayrılabilir: görüntü, dil, planlama disiplinleri gibi. “Yapay zeka zeki makineler, özellikle zeki bilgisayar programları yapma mühendisliği ya da bilimidir.” “Düşünme, anlama, faaliyete geçirmeyi sağlayacak bilgiişleme çalışmasıdır.” “Yapay zeka bağımsız makineler-bu makineler insan olmaksızın karmaşık işler yapabilir-inşa etmek için araştırma yapan bilişsel bilim dalıdır. Bu hedef makinelerin düşünmesini ve anlamasını gerektirir. Bu konuda akıl almaz ilerlemeler sağlanmışsa da hedefe bakıldığında hayal gibi gözükmektedir.”

Yapay zeka, hayatımızın hemen hemen insan hayatının her yerine yerleşmiş durumdadır. Bu sebepten dolayı insan hayatına yön veren, makineler sayesinde üretimin devamlı aktif olmasını sağlayan, insanların çalışma hayatının bir parçası olmakla beraber yapılan işleri kolaylaştırmak insanın çalışma yükünü hafifletmek gibi birçok seçenek düşünülebilir. Günümüzde birçok kurum yapay zeka teknolojilerinden faydalanmaktadır. Gün geçtikçe de bu kurumların yapay zeka teknolojilerinin kullanmasıyla ilerleyen zamanlarda insan yaşam seyri yapay zeka teknolojilerinin kullanılması ile birlikte farklı bir boyutta görülmesi öngörülebilir. Akıllı telefonda kullanılan yapay zeka teknolojileri gibi unsurları barındırdığı için insan hayatında her zaman aktif olarak yerini alan bir teknolojidir.

Yapay Zeka ile Doğal zekayı kıyaslamak istersek;

Artificial Intelligence (Yapay Zeka) ve Natural Intelligence (Doğal Zeka) arasındaki farklar şunlardır:

- **Kendilik:** Doğal zeka, insanlar ve hayvanlar gibi doğal varlıklarda bulunur. Yapay zeka ise, insanlar tarafından yapılan bir teknolojidir.
- **Öğrenme Yöntemi:** Doğal zeka, deneyim ve pratik yoluyla öğrenir. Yapay zeka ise, veri girişi ve programlama yoluyla öğrenir.
- **Üretkenlik:** Doğal zeka, belirli bir noktadan sonra gelişme göstermez. Yapay zeka ise, sürekli veri girişi ve güncelleme ile daha da gelişebilir.
- **Bağımsızlık:** Doğal zeka, insanlar gibi bağımsız değildir. Yapay zeka ise, insanlar tarafından programlanmasına rağmen bağımsız çalışabilir. Bu farklar genel olarak yapay zeka ve doğal zeka arasındaki temel farkları belirtmektedir.
- **Düzenleme:** Doğal zeka, insanlar gibi düzenleme gerektirir ve hata yapabilir. Yapay zeka ise, yapılan programlama doğrultusunda doğru sonuç vermeyi amaçlar, ancak hata yapma ihtimali de vardır.
- **Esneklik:** Doğal zeka, çok sayıda farklı işi yapabilir ve farklı durumlarda esnek tepkiler verebilir. Yapay zeka ise, programlanan işleri yapabilir, ancak farklı durumlar için esnek tepkiler verme kapasitesi sınırlıdır.
- **Duygusal İlişki:** Doğal zeka, duygusal ilişkiler ve hisleri doğal olarak yaşayabilir. Yapay zeka ise, duygusal ilişkiler ve hisleri simüle etme kapasitesi vardır, ancak doğal olarak yaşayamaz.

Bu farklar yapay zekâ ve doğal zekâ arasındaki genel farkları göstermektedir ve her ikisi de kendi alanlarında önemli rol oynamaktadır.

Yapay zeka, bir makine veya yazılım tarafından yapılan, insan beyni gibi düşünen ve çözen aktivitelerdir. Örnekleri sesli komutları anlamaya veya tahmin etmeye çalışan sistemlerdir. Doğal zeka ise, insanlar ve hayvanlar gibi doğal olarak var olan ve insana benzer düşünce ve çözme yetenekleri olan zekâlardır. Yapay zeka ve Doğal zeka, birbirinden farklı şeylerdir ve birbirleriyle karşılaştırılamazlar. Yapay zeka, insana benzer davranışlar sergileyebilir ancak asla doğal davranışların yerini alamaz.

2.2 Yapay Zekada Kullanılan Ajanlar

Yapay zeka ajanı, yapay zeka sistemlerinde kullanılan bir programdır. Bu programlar, belirli bir görev veya amaç için tasarlanmış olup ve belirli bir çalışma alanında hareket etme kabiliyetine sahiptir. Yapay zeka ajanları, verileri toplamak, analiz etmek, karar vermek ve belirli bir amaç için hareket etmek gibi görevleri yerine getirebilir ve bazen insana benzer davranışlar sergileyebilir. Yapay zeka ajanları, yapay zeka uygulamalarında ve hibrid yapay zekâ sistemlerinde kullanılabilir.

Robot bazlı düşünürsek kullanılan ajanlar;

Algılayıcılar: Kamera, sensor, kablosuz bağlantılar

Tepki Vericiler: Çeşitli motorlar, mekanik kollar

Ajan kelimesi yukarı da belirttiğimiz gibi algılayıcılar aracılığıyla bilgileri ilgili ortamlardan toplar ve belli bir zaman sonar girdiye dönüştürür.

Ajanın en önemli işlevi belli başlı algıları ortamlardan topladıktan sonar bu işlevleri harekete dönüştüren yapı diyebiliriz.

- **İnsana Benzeyen Ajanlar:** İnsana benzer davranışlar göstermek için tasarlandı ve kullanılan yapay zeka türüdür.
- **Bölge Ajanları:** Belirli bir alanda hareket etmek için tasarlandı ve kullanılan yapay zeka türüdür.
- **Açık Alan Ajanları:** Geniş ve belirsiz bir alanda hareket etmek için tasarlandı ve kullanılan yapay zeka türüdür.
- **Belge Ajanları:** Verileri toplamak, analiz etmek ve yönetmek için tasarlandı ve kullanılan yapay zeka türüdür.
- **Kavramsal Ajanlar:** Kavramlar arasındaki ilişkileri anlamak için tasarlandı ve kullanılan yapay zeka türüdür.

- **Optimal Ajanlar:** En iyi sonucu elde etmek için tasarlandı ve kullanılan yapay zeka türüdür.

Bu ajanlar, yapay zeka uygulamalarında farklı amaçlar için kullanılabilir ve farklı yapay zeka teknolojilerinin birleştirilmesi ile birlikte kullanılabilir.

2.3 Makine Öğrenmesi Teknikleri

Makine öğrenimi (Machine Learning), yapay zeka (AI) alanında kullanılan bir daldır. Makine öğrenimi, bir bilgisayarın, insan müdahalesi olmadan, verileri analiz etmesini ve bunlardan öğrenmesini sağlayan bir tekniktir. Makine öğrenimi yöntemleri, verileri sınıflandırmak, tahmin etmek, kümelemek veya desenler çıkarmak gibi birçok farklı görevi yerine getirebilirler. Aşağıda, en sık kullanılan makine öğrenimi teknikleri hakkında kısa bir açıklama yer almaktadır:

- **Denetimli Öğrenme (Supervised Learning):** Denetimli öğrenme, belirli bir girdi ile ilişkili belirli bir çıktı verisi kullanarak bir algoritmanın öğrenmesini sağlar. Bu teknik, sınıflandırma ve regresyon gibi görevler için kullanılır.
- **Denetimsiz Öğrenme (Unsupervised Learning):** Denetimsiz öğrenme, girdi verileri arasındaki ilişkileri ortaya çıkarmak için kullanılır. Bu teknik, kümeleme ve boyut azaltma gibi görevler için kullanılır.
- **Yarı Denetimli Öğrenme (Semi-supervised Learning):** Yarı denetimli öğrenme, hem denetimli hem de denetimsiz öğrenme yöntemlerini içeren bir tekniktir. Bu yöntem, az sayıda etiketli veri kullanılarak, daha fazla veri örüntüsü öğrenmeyi sağlar.
- **Takım Öğrenmesi (Ensemble Learning):** Takım öğrenme, birden fazla makine öğrenimi algoritmasını birleştirerek daha yüksek performanslı bir model oluşturmayı sağlar.
- **Derin Öğrenme (Deep Learning):** Derin öğrenme, yapay sinir ağları kullanarak çok katmanlı modeller oluşturur ve karmaşık görevleri gerçekleştirir.

Bu teknik, doğal dil işleme, görüntü tanıma ve ses tanıma gibi alanlarda kullanılır.

Makine öğrenimi teknikleri, birçok farklı görev için kullanılır ve çeşitli sektörlerde yaygın olarak kullanılmaktadır. Yapay zeka'nın alt dalı olan makine öğrenmesini kullanılır. Karmaşık veriden, gürültülü veriden, makine öğrenimi teknikleri kullanılarak sadeleştirilir gerekli istatistiki bilgiler çıkarılır ve sınıflandırmalar yapılır. Makine öğrenmesi teknikleri ile elimizde ki veriyi eğiterek numerik ya da kategorik şekilde sınıflandırabiliriz. Belli aşamaları geliştirdikten sonra seçtiğimiz makine öğrenme teknikleri arasında kuvvetle muhtemel maksimum değeri alan algoritma sınıfını seçeriz ve seçtiğimiz tekniğe göre makine öğrenimi tekniklerinden seçtiğimiz tekniği kullanırsak hata oranı minimum şekilde davranışta bulunur diye yorumlama yapabiliriz. Yapay zekanın bir branşı olan makine öğrenmesi, veriden karmaşık örüntünün tespit edilmesi ve akılcı karar verme için istatistik ve bilgisayarın hesaplama gücünden faydalanır. Makine öğrenme teknikleri sınıflandırma problemlerinde başarılı bir şekilde kullanılmaktadır (Michie, Spiegelhalter ve Taylor: 1994, Sebastiani:2002, Anderson, Michalski ve Mitchell:1983).

Literatürde, saldırı tespit sistemlerinde sıklıkla kullanılan makine öğrenme teknikleri aşağıdaki gibidir.

- NBA - Naive Bayes Algoritması
- SVM - Destek Vektör Makinesi Algoritması
- KNN - K En Yakın Komşu Algoritması
- LRA - Lojistik Regresyon Algoritması
- RFA - Rastgele Orman Algoritması
- DT - Karar Ağaçları Algoritması
- ANN - Yapay Sinir Ağları

2.3.1 NB – Naive Bayes Algoritması

Naive Bayes Algoritması, veri setinde belirli bir olayın gerçekleşme olasılığını belirlemek için kullanılan bir makine öğrenme algoritmasıdır. Algoritma, verilerin belirli bir kategoride nasıl dağıldığını öğrenerek, verileri belirli bir kategoride hangi olasılıkla bulunabileceğini belirler. Naive Bayes Algoritması, Gaussian Naive Bayes,

Multinomial Naive Bayes ve Bernoulli Naive Bayes gibi farklı tip modelleri içermektedir. Her bir model verilerin dağılımını belirlemek için farklı yöntemler kullanır ve her bir modelin farklı veri tipi ve uygulama durumları için uygun olduğu varsayılır. Naive Bayes Algoritması'nın avantajları arasında; hız, kolay uygulanabilirlik, düşük kaynak tüketimi ve verilerin az olduğu durumlarda yüksek performans gibi unsurlar bulunabilir. Dezavantajları arasında ise, veriler arasındaki bağımlılıkları yansıtabilecek "naive" varsayımları, yüksek düzeyde gürültü bulunan veriler ve düşük performans gibi durumlar bulunabilir.

Bayes teoremi, birden fazla etkenin olduğu bir olayın meydana gelmesinde, olayda hangi etkenin payının yüksek olduğunun hesaplanması temeline dayanır. Bayes teoremi aşağıda Eşitlik 1'de gösterildiği gibi ifade edilebilir.

$$P(e|T) = \frac{P(T|e) P(e)}{P(T)}$$

$P(e)$ = e olayının önsel olasılığı

$P(T)$ = T eğitim verisinin önsel olasılığı

$P(T|e)$ = e olayı verildiğinde T'nin koşullu olasılığı

$P(e|T)$ = T eğitim verisi verildiğinde e'nin koşullu olasılığı

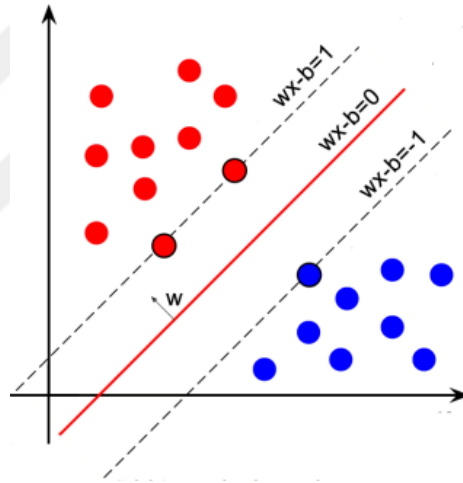
2.3.2 SVM – Destek Vektör Makinesi

SVM (Support Vector Machine), doğal zeka ve makine öğrenmesi alanında kullanılan bir sınıflandırma algoritmasıdır. SVM, veriler arasındaki sınıflandırma hatalarını en aza indirmek için öğrenilen bir hiper düzlem modelini kullanır. Bu model, veriler arasındaki en uygun ayrımı yapar ve verilerin sınıflandırılması için kullanılan "destek vektörleri" belirler. SVM, özellikle büyük veri setlerinde ve karışık veri özellikleri içeren durumlarda oldukça etkilidir ve başka birçok makine öğrenme algoritmasından daha iyi sonuçlar verir. SVM algoritması, verileri iki veya daha fazla sınıfa ayırtırmak için kullanılabilir ve lineer veya daha karmaşık hiper düzlem modelleri kullanabilir. SVM

aynı zamanda regresyon, sıralama ve düzenleme gibi diğer makine öğrenme görevlerini de destekleyebilir.

SVM, çoklu sınıflandırma görevleri, yüksek boyutlu veriler ve özellik sayısının veri sayısından daha fazla olduğu durumlarda özellik seçimi gerektirebilir. SVM aynı zamanda kernel fonksiyonlarını kullanarak verilerin lineer olmayan ayrımlarını tanımlayabilir ve veriler arasındaki non-lineer bağımlılıkları modelleyebilir.

SVM'nin avantajları arasında; etkililik, hassasiyet, hız ve robustluk sayılabilir. SVM'nin dezavantajları arasında ise, hiper parametre seçimi, veri seti boyutu ve yapay veri oluşturma gibi zorluklar bulunabilir.



Şekil 5: Destek vektör makinesi

Kaynak: Sevlı,2021:600

Verinin her bir noktası Eş. 1’de verilen şekilde tanımlanır.

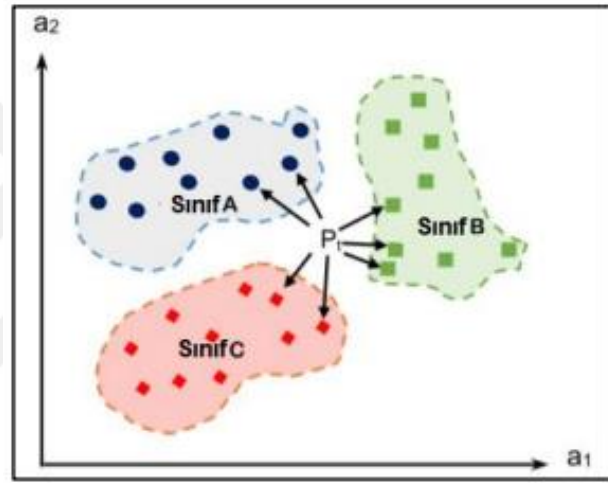
$$\{ (x_i, y_i) \mid x_i \in \mathbb{R}^d, y_i \in \{-1, 1\} \} \quad n \quad i = 1 \quad (1)$$

Formülde x bir girdiyi, y ise -1 ve 1 ile temsil edilen bir sınıfı belirtir. Düzlemde her bir nokta $wx-b$ şeklinde ifade edilir. Burada w düzleme dik olan normal vektörü, b ise kayma miktarıdır. Destek vektör makinesi, karesel optimizasyon yöntemi ile ayırma sınırının bulunmasını sağlar (Shen,2016:61-75)

2.3.3 KNN – K En Yakın Komşu

KNN (K-Nearest Neighbor), verilen bir veri seti üzerinde sınıflandırma yapmak için kullanılan bir yapay zeka algoritmasıdır. Bu algoritmanın temel prensibi, verilen bir örneğin sınıfını belirlemek için veri setindeki en yakın komşularını bulmak ve en çok tekrar eden sınıfı örneğin sınıfı olarak atamaktır. Algoritma, komşuların belirlenmesinde mesafe ölçümü yapması ile çalışır.

K-NN sınıflandırma algoritmasının yapısı şu şekilde örnek verebiliriz. Veri setindeki, veri noktalarından önceden sınıfa dâhil olmuş veri noktalarından, herhangi birisine doğru uzaklık temel alınır sonrasında hedef alınan uzaklığa göre atama yapılır.



Şekil 6: K en yakın komşu

Kaynak: Güvenç ve diğerleri, 2022:1362

K-En Yakın Komşu (KNN) algoritması, sınıflandırma ve regresyon problemlerinde kullanılan bir makine öğrenimi algoritmasıdır. Bu algoritma, veri noktalarını sınıflandırmak veya tahmin etmek için yakındaki K sayıda komşu noktanın etrafındaki etiketi kullanır. KNN algoritması, öncelikle bir uzaklık ölçüsü kullanarak, test örneğinin en yakın K eğitim örneğini belirler. Bu uzaklık ölçüsü, öklid mesafesi, Manhattan mesafesi veya kosinüs benzerliği gibi farklı ölçümlerden oluşabilir. Ardından, KNN algoritması, en yakın K örneğin etiketlerine bakar ve test örneğinin etiketini bu K komşudan en sık tekrarlanan etiketle eşleştirir. Örneğin, bir sınıflandırma problemi ele alalım ve öklid mesafesi ölçüsünü kullanalım. Bir test örneği, 2 boyutlu bir veri noktası

olsun ve $K=5$ olsun. Bu durumda, KNN algoritması, öklid mesafesi ölçüsünü kullanarak eğitim verisindeki tüm örneklerin test örneğine olan mesafesini hesaplar. Sonra, bu örneklerden en yakın olan $K=5$ örnek seçilir. Bu örneklerin etiketleri incelenir ve en sık tekrarlanan etiket, test örneğinin etiketi olarak atanır. KNN algoritması basit, ancak yüksek boyutlu verilerde veya büyük veri kümelerinde hesaplama gücü gerektirir. Ayrıca, veri kümesindeki gürültüye veya açık aykırı değerlere karşı hassastır

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

Veri setinden yeni bir veriyi sınıflandırmak amaçlı olarak algoritmaya verdiğimizde, eğitilmiş veri setinde yer alan k adet en yakın komşuların sınıf etiketleri alınır. Sonraki adımda, ortaya çıkan sınıf etiketlerinden hangisi çoğunluktaysa işleme alınan veri o kümeye dâhil edilir (Güvenç ve diğerleri, 2022:1362).

2.3.4 LR - Lojistik Regresyon Algoritması

Lojistik regresyon (LR), sınıflandırma problemlerinde kullanılan bir makine öğrenimi algoritmasıdır. Bu algoritma, verilerin belirli sınıflar arasında ayrılmasını sağlayan bir çıktı üretir. LR, girdi değişkenleri ve çıktı değişkeni arasındaki ilişkiyi modeller. Ancak, LR, bağımlı değişkenin kategorik olduğu sınıflandırma problemlerinde kullanılır. LR, çıktı olarak bir olasılık değeri verir ve bu olasılık değeri, sınıflandırma yapmak için bir eşik değeri ile karşılaştırılır. Örneğin, bir sınıflandırma probleminde, bir öğrencinin sınav sonucunu alın ve öğrencinin sınavı geçip geçmediğini belirleyin. Sınav sonucu, girdi değişkeni olarak kullanılabilir ve çıktı değişkeni, öğrencinin sınavı geçip geçmediği (1 veya 0) olabilir. LR, sınav sonucu ile öğrencinin sınavı geçme olasılığı arasındaki ilişkiyi modeller. Sonuç olarak, bir öğrencinin sınavı geçme olasılığı elde edilir ve eşik değeri kullanılarak, öğrencinin sınavı geçip geçmediği belirlenir. çıktı değişkeni arasındaki ilişkiyi modellemek için bir sigmoid fonksiyonu kullanır. Bu fonksiyon, bir olasılık değeri üretir ve bu olasılık değeri, sınıflandırma yapmak için bir eşik değeri ile karşılaştırılır. Sigmoid fonksiyonu, girdi değişkenlerinin doğrusal bir kombinasyonu ile hesaplanır. Bu doğrusal kombinasyon, LR modelinde öğrenilen katsayılar ve girdi değişkenlerinin değerlerine dayanır. LR algoritması, girdi değişkenlerinin normal dağılım göstermesi ve birbirleriyle doğrusal olarak bağlantılı olmaması gibi bazı varsayımlara dayanır. Ayrıca, LR, veri kümesindeki gürültüye ve

aykırı değerlere karşı hassastır. LR, sınıflandırma problemlerinde kullanılan en yaygın algoritmalarından biridir ve özellikle tıp, ekonomi, sosyal bilimler ve mühendislik alanlarında kullanılır. LR, özellikle iki sınıflı sınıflandırma problemleri için uygundur, ancak çoklu sınıf sınıflandırması için de kullanılabilir. Denklem (3)'deki gibi tanımlanmış çoklu doğrusal regresyon fonksiyonunu kullanır.

$$\text{logit}(p)=\log (p(y=1) 1-(p=1)) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \dots + \beta_kX_k \quad (3)$$

Çıkan denklem sonucu p, karakteristik şeklinin olma ihtimali vardır. Bir diğer önemli değer de odds değeri olarak adlandırılan değerdir. Denklem (3) ile verilen bu değer karakteristik özelliğinin var olma olasılığı (p) değerinin, karakteristik özelliğinin var olmama olasılığı (p-1) değerine bölünmesiyle ortaya çıkan bir değerdir.

$$\text{odds} = \frac{P}{P-1} = \frac{\text{karakteristiğinin var olma olasılığı}}{\text{karakteristiğinin var olmama olasılığı}} \quad (4)$$

2.3.5 RF - Rastgele Orman Algoritması

Rastgele ormanlar veya rastgele karar ormanları, eğitim zamanında çok sayıda karar ağacı oluşturarak çalışan, sınıflandırma, regresyon ve diğer görevler için bir topluluk öğrenme yöntemidir. Sınıflandırma görevleri için, rastgele ormanın çıktısı, çoğu ağaç tarafından seçilen sınıftır. Regresyon görevleri için, tek tek ağaçların ortalama veya ortalama tahmini döndürülür (Ho, 1995: 278-282). Rastgele orman algoritması genellikle karar ağaçlarından daha iyi performans gösterir, ancak doğruluk değerleri gradyan destekli ağaçlardan daha düşüktür. Ancak, veri setinde yer alan verilerin özellikleri performanslarını etkileyebilir (Piryonesi ve Diraby,2021:1-17).

Rastgele orman algoritmasında yer alan ağaçların her birinin öğrenmesi aşamasında kullanılan genel teknik bootstrap veya bagging yöntemidir. Bu yöntemin işleyişinde, eğitim sonrası, görünmeyen örnekler x' için tahminler, x' üzerindeki tüm bireysel regresyon ağaçlarından gelen tahminlerin ortalaması alınarak aşağıdaki formül ile veya sınıflandırma ağaçları söz konusu olduğunda oy çoğunluğuna bakılarak yapılabilir (James ve diğerleri,2013:316-321).

$$f = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (5)$$

Bu bootstrap prosedürü, bias'ı artırmadan modelin varyansını azalttığı için daha iyi bir performans elde edebilmeyi sağlamaktadır (James ve diğerleri,2013:316-321).

2.3.6 ANN - Yapay Sinir Ağları

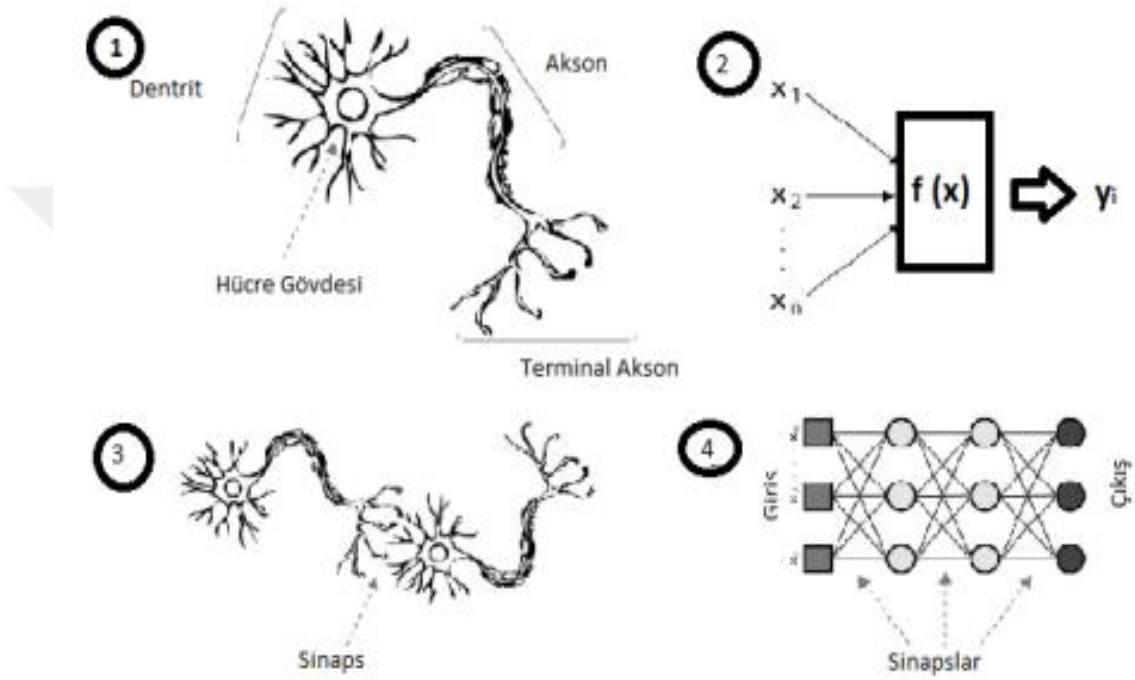
Yapay sinir ağları YSA, insan beynine benzetildiği için üstün özellikler atfedilmiştir. İnsan beyni de sinirlerden oluştuğu için ve sinirlerin içinden geçen nöronlar olduğu için YSA bu modelle çalıştırması hedeflenmiştir. İnsanda bulunan sinir hücreleri sayesinde makine adı altında değerlendirsek sensörlere benzetebiliriz. Dokunduğunda, hissettiğinde, çarptığında sinir hücresi beyne anında sinyal gönderiyor. YSA da insana özgü objeleri nitelendirip modellendiriliyor. YSA nöronları var ve bu nöronlar karar vermeye yarıyorlar. Bir nöron kendi üzerine gelen sinyalleri topluyor ve bu sinyallere göre bir aksiyon alıyor. Çalışma yapısını da şöyle açıklayabiliriz:

Nörona her yerden bilgiler geliyor, bu akan bilgilere göre nöron karar veriyor. Bu toplanan veriler belli bir threshold value'nin yani belli bir değere gelince 1 ateşleniyor, belli bir değerin altında olunca da 0 olarak ateşleniyor. Axon üzerinden yani bir sinir hücresi üzerinden sinyali iletiyor ya da iletmiyor.

Yapay bir sinir hücresi, biyolojik bir sinir hücresinin temel davranışlarından esinlenen matematiksel modeli ortaya koyan bir algoritma veya fiziksel araç olarak tanımlanabilir. Biyolojik sinir hücresinin tanımından hareket ederek, yapay bir sinir hücresinin diğer sinir hücrelerinden aldığı sinyalleri bünyesinde topladığını ve toplam sinyal birikiminin belli bir eşiği aştığı anda, bu yapay sinir hücresinin kendi sinyalini bir başka sinir hücresine ilettiği söylenebilir (Ataseven, 2013). Biyolojik sinir ağlarının sinir hücreleri olduğu gibi yapay sinir ağlarının da yapay sinir hücreleri vardır. Yapay sinir hücreleri mühendislik biliminde proses elamanları olarak da adlandırılmaktadır. her proses elemanın 5 temel elemanı vardır. Bunlar; girdiler, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktılarıdır (Öztemel,2006)

Yapay Sinir Ağları (YSA), insan beyninin çalışma prensiplerinden esinlenilerek oluşturulmuş bir makine öğrenimi teknolojisidir. YSA'lar, bilgisayarların karmaşık görevleri yerine getirebilmesi için tasarlanmıştır. YSA'lar, verileri analiz etmek, sınıflandırmak, örüntüleri tanımak ve tahmin yapmak için kullanılırlar. Yapay sinir ağları, insan beynindeki nöronların çalışma prensiplerinden esinlenilerek tasarlanmıştır.

Nöronlar, beyindeki sinir hücreleridir ve birbirleriyle bağlantılıdır. Benzer şekilde, yapay sinir ağları da birbirleriyle bağlantılı olan bir dizi matematiksel işlem düğümüdür. YSA, girdi katmanı, saklı katmanlar ve çıktı katmanı olmak üzere üç katmandan oluşur. Girdi katmanı, verilerin alındığı katmandır. Saklı katmanlar, verilerin işlenerek öğrenildiği katmanlardır. Çıktı katmanı ise sonuçların verildiği katmandır.



Şekil 7: Biyolojik sinir hücresi ve yapay sinir ağı

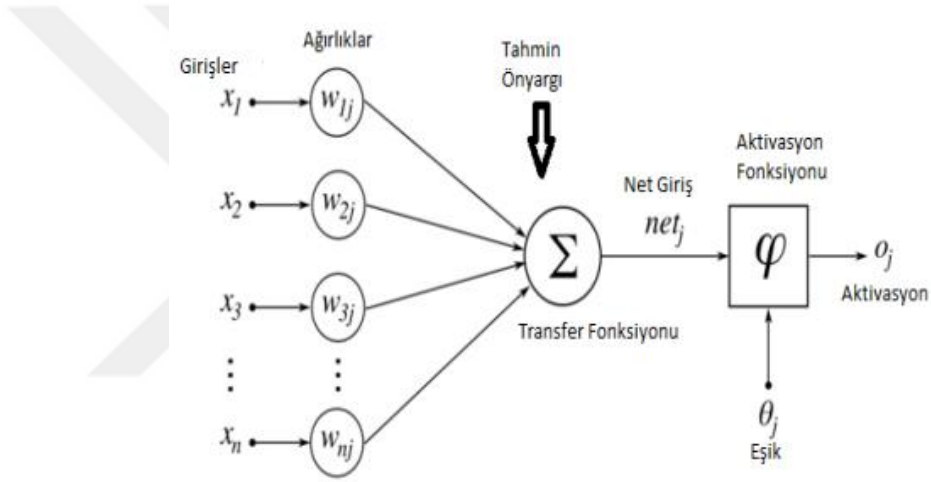
Kaynak: Matarollo, 2013

İnsana ait biyolojik özelliklerin, YSA karşılık gelen benzetimli halini görüyoruz.

İnsana ait biyolojik sinir sistemi elemanlarıyla makine öğrenimi dalından birisi olan yapay sinir sistemine karşılık neler geldiğini tablo 1 de görmekteyiz.

Tablo 1**Biyolojik sinir sistemi elemanları ve yapay sinir sisteminde karşılıkları**

Biyolojik Sinir Sistemi	Yapay Sinir Sistemi
Nöron	İstemci elemanı
Dentrit	Toplama fonksiyonu
Hücre gövdesi	Transfer fonksiyonu
Aksonlar	Yapay nöron çıkışı
Sinapslar	Ağırlıklar

**Şekil 8:** Yapay sinir hücresi**Kaynak:** URL 1, 2018

1'den başlayıp, n'ye kadar nöronlar mevcuttur. Girişler ise X_1 'den başlayıp, X_n 'e kadar devam etmiştir. Giriş alanına sensor ya da duyarga da denilebilir. X_1 den W_1 'e giden iletkeni ya da hücreyi takip edersek o arada bir direnç oluşuyor. Burada bir weight – ağırlık oluşuyor. Duyargalar ile nöronlar arasındaki her çizgi de bir ağırlık değeri var. 1'den, n'ye kadar olan input'tan ya da duyargadan her biri farklı ağırlık değeri ile çarpılıyor ve bu çarpılan ağırlık değerlerine göre Transfer Fonksiyonu alanında değerlerin hepsi toplanıyor. Toplam değer threshold value'yi geçiyorsa transfer fonksiyonundan ateşlenip bir sonraki sürece giriyor.

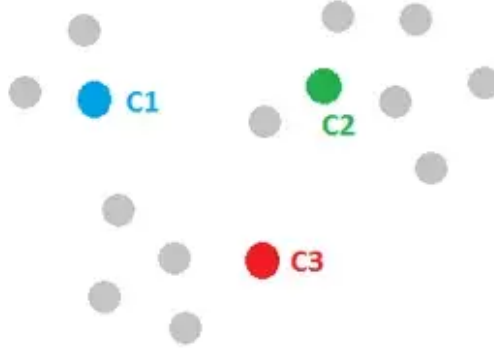
Aktivasyon fonksiyondan bahsederek input değerinden gelen değerler negatif ise aktivasyon fonksiyonundan eksi olarak atışılacak, input değerinden gelen değerler pozitif ise artı değer atışılacak.

2.4 Makina Öğrenme Algoritmaları Kümeleme

Makine öğrenimini iki başlık altında değerlendirdiğimiz de denetimli öğrenme yapısında sınıflandırma algoritmalarının hangileri olduğuna değindik. Kümeleme algoritmalarından bahsederek; Kümeleme algoritması denetimsiz öğrenme kategorisi arasında yer alır. Bir modelde gruplara ayırdığımızda hangi grupta kalacağını kümeleme algoritmasını kullanarak belirlenir. Herhangi bir denetim mekanizması olmadan denetimsiz bir şekilde gruplara ayrılır. Oluşan modelin ne şekilde davranış sergileyeceği işlenecek olan algoritmaya bağlıdır. Grupta oluşan verinin türüne bakılır. Veri kendine ait özellikler bilinmiyorsa etiketsiz verilerdir. Etiketsiz verilerin özellikleri bilinmediği için gruplandırma yapılır bu durum da kümeleme yöntemi kullanılmış olur. Eğer verilerin özellikleri biliniyor olsaydı, etiketli veri olmuş olurdu bu durumda da sınıflandırma algoritmaları kullanılacaktı. Kümeleme tekniğini incelediğimiz de en çok kullanılan algoritmalar arasında K-Means algoritması ve Hierarchical kümeleme algoritmasını göreceğiz. K-Means algoritması ile gruplara ayıracağız. Gruplara ayırmak için benzer özelliklerine göre ayırma tekniği uygulanması ön görülmektedir. Belirli bir veri kümesindeki öğeleri benzerlik ölçüleri kullanarak kümelere ayırır. Bu kümeleme işlemi, benzer özelliklere sahip öğeleri bir arada gruplandırarak veri setindeki yapılardan farklılıkları ortaya koyar. K-Means, verileri belirli sayıda küme oluşturmak için kullanılan bir yöntemdir. Bu yöntemde, başlangıçta rastgele seçilen merkezler etrafında benzer veriler bir araya getirilir ve her adımda merkezler güncellenerek gruplamalar iyileştirilir.

2.4.1 K-Means Algoritması

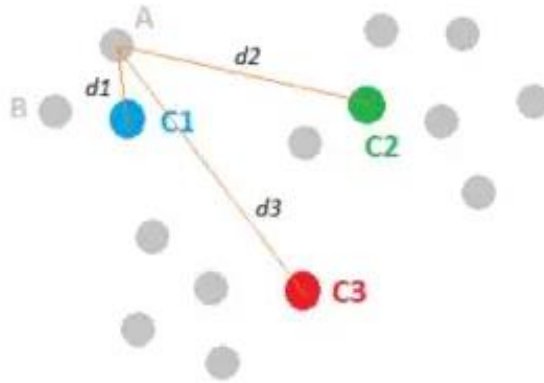
Belirlenen modele göre oluşan kümeler için rastgele merkezde noktası seçilir.



Şekil 9: K-Means algoritması – Rastgele nokta seçilmesi

Kaynak: Harman, 2020

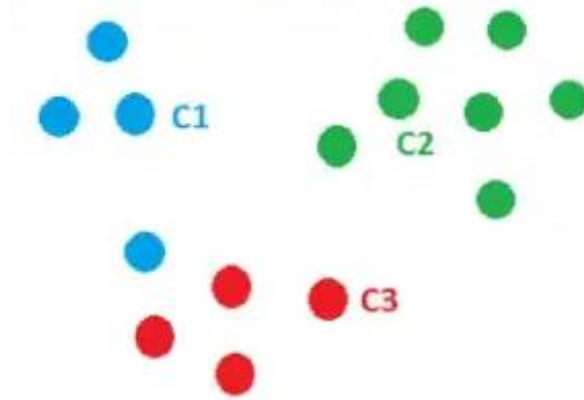
Merkezdeki elaman haricindeki elemanlar kendisine en yakın merkez noktasına göre kümelenerek yerlerini alır. Burada bulunan her elemanın merkez noktalarına göre uzaklıkları ölçülerek bulunur.



Şekil 10: K-Means algoritması – Merkez noktasına göre uzaklık

Kaynak: Harman, 2020

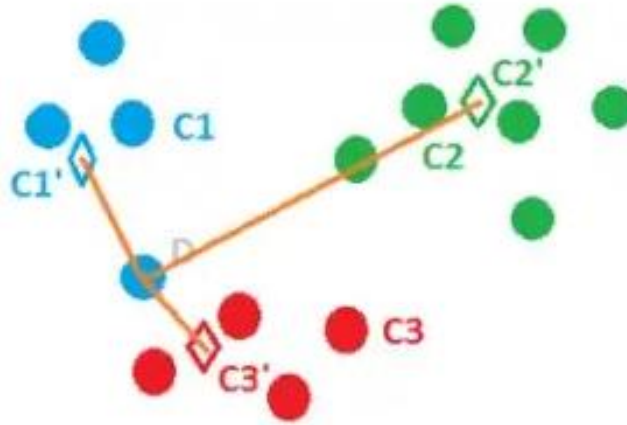
Kümelenen elemanları incelediğimizde bu şekilde yer aldığını görmekteyiz



Şekil 11: K-Means algoritması – Kümelenen elemanlar

Kaynak: Harman, 2020

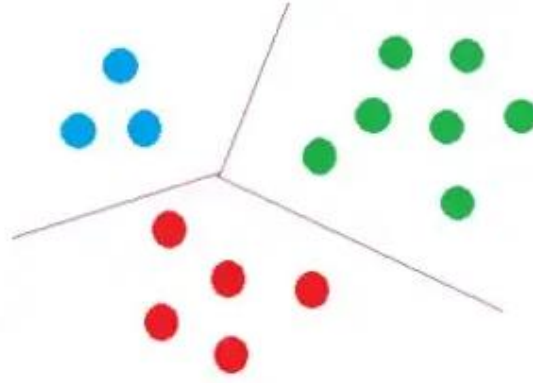
Artık bu kümeleme modeline göre yeniden merkez noktası hesaplaması yapılması ön görülür.



Şekil 12: K-Means algoritması – Yeni kümeleme modeli

Kaynak: Harman, 2020

Sonrasında oluşan kümeleme ile merkez noktası stabil duruma gelene kadar yapılan baştaki iki adım yeniden tekrarlanarak yapılır.



Şekil 13: K-Means algoritması – Tekrar edilen adımlar

Kaynak: Harman, 2020

K-Means algoritmasının avantajlı olduğu gibi büyük veri, küçük veri olarak incelediğimizde ideal algoritma yapısını işleyebiliyoruz. Dezavantajından bahsetmek gerekirse problem oluşturma ihtimali olan süreç kaç adet kullanıcı olma ihtimali olan kullanan kişiden alındığı için belirlenen her sayı , her zaman normal şartlarda optimum değeri verme ihtimali olmayabilir. Bu birinci problem olarak algılanabilir. Diğer problem ise kullanıcıdan atanan merkez noktaları devamlı olarak başarılı olmayabilir. Nedeni ise algoritma devamlı olarak kendini yenilemekte yani algoritma başlangıçta ilk halinde rastgele atananlara göre kendi merkezini oluşturdu. Buna göre şekillenmektedir. Optimum küme belirlemek için WCSS (within-cluster sums of squares)adındaki yöntemi uyguluyoruz. WCSS yönetimi içi aşağıdaki formül kullanılır:

$$WCSS = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

Şekil 14: WCSS denklemi

2.5 Doğal Dil İşleme Nedir, Neden Önemlidir.

Doğal dil işleme (NLP), bilgisayarların ve yazılımların insan dilini anlama ve kullanma yeteneğidir. Bu, metinleri okuma, anlama, yazma ve konuşma gibi insanların yaptığı

işlemleri gerçekleştirmek için gereken teknolojileri içerir. NLP, birçok alanda önemli rol oynar, örneğin arama motorları, sosyal medya, makine öğrenimi, müşteri hizmetleri ve sağlık hizmetleri gibi. NLP, özellikle veri madenciliği ve makine öğrenimi gibi alanlarda çok yararlıdır. Örneğin, NLP teknolojisi sayesinde, metinler içinde anahtar kelimeler ve cümleleri bulmak, metinleri kategorize etmek ve özetlemek mümkün hale gelir. Bu, veri madenciliği için önemli bir araçtır çünkü veri madenciliği sürecinde, veri setlerinde anlamlı bilgiye ulaşmak için verilerin işlenmesi gerekir. NLP, ayrıca makine öğrenimi sistemlerinin daha etkili bir şekilde çalışmasını sağlar. Örneğin, NLP teknolojisi sayesinde, metin verileri sisteme daha kolayca yüklenebilir ve sistem, metinde yer alan anlamlı bilgileri daha kolayca anlayabilir. Sonuç olarak, NLP önemli bir teknolojidir çünkü insan dilini anlama ve kullanma yeteneği, birçok alanda kritik öneme sahiptir. Bu teknoloji sayesinde, veri madenciliği ve makine öğrenimi gibi alanlarda daha etkili sonuçlar elde edilir.

NLP, ayrıca önemli bir rol oynar diğer alanlarda da. Örneğin, sosyal medya ve müşteri hizmetleri gibi alanlarda, NLP sayesinde insanların yazdığı metinleri anlamak ve yanıtlamak daha kolay hale gelir. Bu, müşteri hizmetleri gibi alanlarda daha hızlı ve etkili bir hizmet sunmak için önemlidir. Doğal dil işleme yapısını artık çoğu kurum kendi bünyelerinde kullanmaktadır. Özellikle hakaret içerikli cümle olsun, saldırgan cümleler olsun bu tür cümleler resim formatında olduğunu varsayalım. Kullanılan algoritmalar sayesinde resimden, text'e dönüştürüp, yazılan cümleleri algılayabiliyor.

2.6 Doğal Dil İşlemenin Kullanım Alanları Ve Analiz Edilmesi

Doğal Dil İşleme (DDI), insan dilinin anlaşılması, analizi ve senteziyle ilgilenen bir disiplindir. DDI, yazı veya konuşma gibi doğal dil verilerinin işlenmesi ve anlamlı bilgiye dönüştürülmesiyle ilgilidir. DDI, insan dilinin bilgisayarlar tarafından anlaşılabilmesi için kullanılır ve birçok farklı endüstride kullanım alanı bulunur. DDI verileri analiz etmek için birçok farklı teknik kullanır. Bu teknikler arasında metin madenciliği, doğal dil işleme araçları, yapay sinir ağları, derin öğrenme, doğrusal regresyon, karar ağaçları, destek vektör makineleri gibi teknikler yer alır. Bu teknikler, verilerin işlenmesini, temizlenmesini, özelliklerinin çıkarılmasını, öznetelikleri

kodlamayı ve sınıflandırmayı kolaylaştıran birçok farklı algoritma ve yöntem içerir. DDI teknikleri, verilerin dilbilgisel yapılarını çözmek için morfolojik analiz, sentaks analizi ve anlamsal analiz gibi farklı yöntemler kullanır. Bu teknikler, bir metnin anlamını ve içeriğini anlamak için kullanılır. Morfolojik analiz, kelimenin yapısal özelliklerini ve köklerini belirler. Sentaks analizi, kelimelerin bir araya gelmesiyle oluşan cümlelerin yapısını belirler. Anlamsal analiz, metnin anlamını anlamak için kullanılır ve kelimelerin anlamlarını ve bağlantılarını belirler.

Doğal dil işleme (NLP), birçok farklı alanda kullanılır ve analiz edilir.

Arama Motorları: NLP, arama motorlarında metinler içinde anahtar kelimeler ve cümleleri bulmak ve metinleri kategorize etmek için kullanılır.

Sosyal Medya: NLP, sosyal medya platformlarında yazılan metinleri anlamak ve yanıtlamak için kullanılır. Sosyal medya platformlarında kullanıcılar, metin tabanlı içeriklerini paylaşırlar ve bu içerikler NLP algoritmaları tarafından işlenebilir. NLP, sosyal medyada birçok farklı uygulamada kullanılabilir. İşte bazı örnekler:

- Duygu analizi: Sosyal medya platformlarında kullanıcılar, duygusal içeriklerini paylaşırlar.
- Trend analizi: Sosyal medya platformları, kullanıcıların popüler konular hakkında konuşmalarını sağlar. Bu konuşmalar, NLP algoritmaları kullanılarak analiz edilerek trendler belirlenebilir ve bu trendlere göre işletmelerin, pazarlama stratejilerini belirlemesi mümkün olabilir.
- Marka yönetimi: Markalar, sosyal medya platformlarında müşterileriyle etkileşim kurarlar ve müşterilerinin marka hakkındaki düşüncelerini paylaşırlar.

Makine Öğrenimi: NLP, makine öğrenimi sistemlerinin daha etkili bir şekilde çalışmasını sağlar. Metin verileri sisteme daha kolayca yüklenebilir ve sistem, metinde yer alan anlamlı bilgileri daha kolayca anlayabilir. Makine öğrenimi, bir veri kümesindeki kalıpları tanımlayarak öğrenir. NLP'de, model genellikle bir dizi kelime veya metin parçası olarak temsil edilen bir dizi belgeden oluşur. Makine öğrenimi teknikleri, NLP'de yaygın olarak kullanılan birkaç görevi içerir:

- Sınıflandırma: Belirli bir belgenin hangi kategoriye ait olduğunu belirlemek için kullanılır.
- Etiketleme: Belirli bir belgedeki kelimelerin türlerini (isim, sıfat, fiil vb.) belirlemek için kullanılır.
- Şablon eşleştirme: Belirli bir deseni içeren belgeleri bulmak için kullanılır.
- Duygu analizi: Bir belgenin duygusal tonunu belirlemek için kullanılır.
- Dil çevirisi: Bir dilden diğerine belge çevirisi yapmak için kullanılır.

Makine öğrenimi teknikleri, NLP'deki bu görevleri gerçekleştirmek için kullanılır ve doğal dildeki metinleri işlemenin daha hızlı ve daha doğru bir yolunu sağlar.

Sağlık Hizmetleri: NLP, sağlık hizmetleri alanında, doktorlar ve diğer sağlık profesyonelleri, hasta kayıtlarındaki bilgileri daha hızlı ve doğru bir şekilde okuyabilir ve anlayabilir.

Sesli Asistanlar ve Sesli Aramalar: NLP, insanların sesli komutlarını anlamak ve yerine getirmek için sesli asistanlar veya sesli arama uygulamaları gibi cihazlar geliştirilir.

Analiz edilmesi için, NLP teknolojisi sayesinde metinler içinde anahtar kelimeler ve cümleler bulunur, metinler kategorize edilir, özetleri yapılır, duygu analizi yapılır, çeviri yapılır ve dil öğrenimi yapılır. Bu analizler, arama motorları, sosyal medya, makine öğrenimi, sağlık hizmetleri ve sesli asistanlar ve sesli aramalar gibi alanlarda kullanılır.

Bazı çalışma alanları kapsamında OCR işlemlerinde kullanılan algoritmalar sayesinde hakaret içerikli cümleleri, saldırgan tavrılı cümleleri anında doğal dil işleme işlemleri ile ilgili algoritmalarla işleyip, kullanılan algoritmalarından hangisi yüksek ise diğer veri setlerinde bu algoritma kullanılıyor. Bu işlemler sayesinde resimde yani text olmayan alanda mevcut kelime, cümle yapısında OCR, NLP teknolojileri ile anında text'e dönüştürüp, NLP de kullanılan algoritmaya göre geliştirme yapılabilir. NLP sayesinde insanların sosyal medya platformlarında paylaştığı cümlelere göre duygu analizi yapıp, insanların ruh hali analiz edilebilir.

BÖLÜM 3: SOSYAL MEDYA PLATFORMLARI VE TARİHÇESİ

Sosyal medya, internet kullanıcılarının birbirleriyle iletişim kurabilecekleri ve içerik paylaşabilecekleri platformlardır. Bu platformlar, kullanıcıların fotoğraflar, videolar, metinler ve diğer içerikleri paylaşmasına, yorum yapmasına ve beğenmesine olanak tanır.

Sosyal medya platformları, ilk olarak 1999 yılında Six Degrees adlı bir sosyal ağ oluşturuldu. Ancak, sosyal medya platformları, gerçek anlamda popüler hale 2003 yılında MySpace adlı bir sosyal ağ ile geldi. Daha sonra, Facebook (2004) ve Twitter (2006) gibi platformlar popüler hale geldi. Bu platformlar, kullanıcıların arkadaşları ve aileleriyle iletişim kurmasına olanak tanıdı ve sosyal medya, günlük hayatta kullanılan bir araç haline geldi.

2010 yılından itibaren, Instagram, Pinterest, Snapchat ve TikTok gibi platformlar ortaya çıktı. Bu platformlar, kullanıcıların fotoğraflar ve videolar paylaşmasına odaklandı ve kullanıcıların kendilerini ifade etmelerine ve içeriklerini keşfetmelerine olanak tanıdı.

Son yıllarda, sosyal medya platformları, işletmeler ve markalar tarafından da aktif olarak kullanılmaya başlandı. Bu platformlar, işletmelerin ve markaların müşterileriyle iletişim kurmasına, reklamlarını yayınlamasına ve müşteri yönetimine olanak tanır.

Sosyal medya platformları, günümüzde hayatımızın önemli bir parçası haline geldi ve kullanıcıların arkadaşları ve aileleriyle iletişim kurmasının yanı sıra, işletmeler ve markalar için de önemli bir pazarlama aracı haline geldi.

3.1 Sosyal Medya Kavramları Ve Tarihçesi

Sosyal medya, internet üzerinden insanların bir araya gelerek iletişim kurabildiği, paylaşım yaptığı ve içerik oluşturabildiği platformlar olarak tanımlanabilir. Örnek olarak Facebook, Twitter, Instagram, TikTok gibi uygulamalar sosyal medya platformlarıdır. Bu platformlar, kullanıcıların fikirlerini, düşüncelerini, fotoğraflarını, videolarını ve diğer içeriklerini paylaşmasına olanak tanır.

Sosyal medya, aynı zamanda reklam ve pazarlama amaçları için de kullanılabilir. Örneğin, bir şirket sosyal medya hesapları aracılığıyla ürünlerini ve hizmetlerini

tanıtabilir ve müşterileriyle iletişim kurabilir. Ayrıca, sosyal medya, sosyal ağlar aracılığıyla insanların birbirleriyle iletişim kurmasına ve arkadaşlıklar kurmasına olanak tanır.

Sosyal medya kullanımının yaygınlığı, günümüzde insanların hayatlarının önemli bir parçası haline gelmiştir. Sosyal medya kullanıcıları, haberleri, bilgileri ve eğlence içeriğini takip edebilir, kendilerine ilgi duydukları konular hakkında bilgi edinebilir ve diğer kullanıcılarla etkileşimde bulunabilirler.

3.2 Sosyal Medyada En Çok Kullanılan İçerik Paylaşım Siteleri:

En çok kullanılan sosyal medya içerik paylaşım siteleri arasında kısa olarak bahsedersek:

- **Facebook:** Dünya çapında en yaygın kullanılan sosyal medya platformu olarak kabul edilir. Kullanıcılar, arkadaşları ve aileleriyle iletişim kurabilir, paylaşımlar yapabilir ve gruplar oluşturabilir.
- **YouTube:** Video paylaşım platformudur. Kullanıcılar, kendi yaptıkları videoları yükleyebilir ve diğerlerinin videolarını izleyebilirler.
- **WhatsApp:** Sosyal medya aracılığıyla mesajlaşma, sesli ve görüntülü arama yapabilen bir instant messaging uygulamasıdır.
- **Instagram:** Fotoğraf ve video paylaşımı için kullanılan bir sosyal medya platformudur. Kullanıcılar, fotoğraflarını ve videolarını paylaşabilir, diğer kullanıcıların içeriğine erişebilir ve yorum yapabilirler.
- **Twitter:** Kısa mesajlar olarak bilinen "tweet"ler aracılığıyla haberleri, düşünceleri ve bilgileri paylaşan bir sosyal medya platformudur.
- **TikTok:** Kısa formatlı videolar paylaşım platformudur. Özellikle genç kullanıcılar arasında popüler olan uygulama, dans, müzik, komedi gibi birçok konuda içerikler sunmaktadır.

3.2.1 Youtube

YouTube, Google tarafından 2005 yılında satın alınan bir video paylaşım platformudur. Kullanıcılar, kendi yaptıkları videoları yükleyebilir ve diğer kullanıcıların videolarını izleyebilirler. Platform, ücretsiz ve açık kaynaklıdır ve herkesin kullanabildiği bir hizmettir.

YouTube, kullanıcıların video içeriği oluşturmaya ve yaymasına olanak tanır. Bu içerikler arasında müzik videoları, belgeseller, komedi videoları, dersler ve eğitim videoları, haberler gibi birçok farklı konuda video bulunur. Kullanıcılar ayrıca video içeriklerini beğenerek, yoruma ve paylaşarak etkileşimde bulunabilir. Ayrıca YouTube, birçok ünlü ve popüler YouTuber'ların yer aldığı bir platformdur.

3.2.2 İnstagram

İnstagram, Facebook tarafından 2012 yılında satın alınan bir fotoğraf ve video paylaşım platformudur. Kullanıcılar, kendi çektikleri fotoğrafları ve videoları paylaşabilir, diğer kullanıcıların içeriğine erişebilir ve yorum yapabilirler. Ayrıca, kullanıcılar arasında arkadaşlık ve takip ilişkileri oluşabilir.

İnstagram, kullanıcıların fotoğraf ve video içeriği oluşturmaya ve yaymasına olanak tanır. Bu içerikler arasında kişisel fotoğraflar, moda, yemek, tatil, doğa gibi birçok konuda fotoğraf ve video bulunur. Kullanıcılar ayrıca fotoğraflar ve videoları "beğenme" ve yoruma göndererek etkileşimde bulunabilir. Instagram, ayrıca birçok ünlü ve popüler kullanıcıların yer aldığı bir platformdur.

3.2.2 Sosyal Medyada En Çok Kullanılan Sosyal Ağlar:

3.2.2.1 Twitter

Twitter, 2006 yılında kurulmuş bir sosyal ağ platformudur. Kullanıcılar, kısa mesajlar olarak bilinen "tweet"ler aracılığıyla haberleri, düşünceleri ve bilgileri paylaşabilirler. Tweetler genellikle 280 karakter veya daha az süren kısa metinlerdir. Twitter, kullanıcıların diğer kullanıcıları takip etmelerine ve kendilerine takip edilmelerine olanak tanır. Bu sayede kullanıcılar, arkadaşları, aileleri, ünlüler veya ilgi duydukları kişi veya kurumların güncel haberlerini ve düşüncelerini takip edebilirler. Ayrıca, Twitter, haber ve bilgi almak için popüler bir platformdur, çünkü kullanıcılar, güncel haberleri ve gelişmeleri hızlı bir şekilde öğrenebilirler.

Twitter ayrıca etkileşimli bir platformdur. Kullanıcılar, tweetleri "beğenme", yorum yapma veya retweet yapma (paylaşma) gibi şekillerde etkileşimde bulunabilirler. Ayrıca, Twitter, hashtag'ler (#) aracılığıyla konuları ve etiketleri izlemeye olanak tanır. Bu sayede kullanıcılar, belirli bir konuda ilgi duydukları tweetleri ve kullanıcıları bulabilirler. Twitter'ın popülaritesi, özellikle siyasi ve sosyal konuların tartışıldığı, haberlerin ve bilgilerin paylaşıldığı veya ünlü kişilerin paylaştığı bilgileri takip etmek için kullanılması sebebiyle artmıştır. Ayrıca, birçok kurum ve şirket de Twitter aracılığıyla müşteri hizmetleri, pazarlama veya reklam faaliyetleri yürütmektedir.

3.2.2.2 LinkedIn

LinkedIn, 2002 yılında kurulmuş bir sosyal ağ platformudur. İşletmeler ve profesyoneller arasında bağlantı kurmalarını ve iş fırsatlarını paylaşmalarını amaçlar. Kullanıcılar, kendi profillerini oluşturabilirler ve bu profillerde eğitim, iş deneyimi, beceriler gibi bilgileri paylaşabilirler. Ayrıca, kullanıcılar arasında arkadaşlık ve iş ilişkileri oluşabilir.

LinkedIn, iş arama ve iş fırsatlarının paylaşılması için popüler bir platformdur. Kullanıcılar, iş arama veya iş değiştirme amacıyla kullanabilirler veya işletmeler için potansiyel çalışanları bulmak için kullanabilirler. Ayrıca, LinkedIn, profesyoneller arasında bilgi paylaşımı ve network oluşturma için de kullanılır. Örneğin, bir profesyonel, sektöründe uzman olan kişilerle iletişim kurabilir veya kendini tanıtabilir. LinkedIn, aynı zamanda bir işletmeler platformudur. İşletmeler, kendi sayfalarını oluşturabilir ve ürün ve hizmetlerini tanıtabilirler. Ayrıca, işletmeler, çalışanlarının profillerini yayınlatabilir ve potansiyel çalışanlarla iletişim kurabilirler.

LinkedIn, ayrıca bir eğitim platformudur. Kullanıcılar, eğitimlerini ve sertifikalarını profillerinde yayınlatabilirler veya eğitim kurumları tarafından sunulan eğitimleri keşfedebilirler. LinkedIn, ayrıca bir gruplar platformudur. Kullanıcılar, belirli bir konuda ilgi duydukları grupları katılabilir veya kendi gruplarını oluşturabilirler. Bu sayede, kullanıcılar, benzer ilgi alanlarına sahip kişilerle iletişim kurabilir ve bilgi paylaşabilirler.

3.2.2.3 Facebook

Facebook, 2004 yılında Mark Zuckerberg tarafından kurulmuş bir sosyal ağ platformudur. Kullanıcılar, kendi profillerini oluşturabilir ve bu profillerde kişisel

bilgileri, fotoğrafları, videoları ve güncellemeleri paylaşabilirler. Ayrıca kullanıcılar arasında arkadaşlık ve iletişim ilişkileri oluşabilir. Facebook, kullanıcıların arkadaşları ve aileleriyle iletişim kurmasına olanak tanır. Kullanıcılar, mesajlaşma, sesli ve görüntülü arama gibi özellikleri kullanarak iletişim kurabilirler. Ayrıca, Facebook, kullanıcıların paylaşımlar yapmasına ve gruplar oluşturmaya olanak tanır. Bu sayede kullanıcılar, arkadaşları ve aileleriyle birlikte etkinlikler düzenleyebilir veya ilgi duydukları konularda gruplar oluşturabilirler.

Facebook, ayrıca bir reklam ve pazarlama platformudur. İşletmeler, Facebook aracılığıyla ürünlerini ve hizmetlerini tanıtabilir ve müşterileriyle iletişim kurabilirler. Ayrıca, Facebook, birçok ünlü ve popüler kullanıcıların yer aldığı bir platformdur. Facebook, ayrıca bir eğlence platformudur. Kullanıcılar, oyunlar oynayabilir veya eğlence içerikleri izleyebilirler. Ayrıca, Facebook, birçok kullanıcının yer aldığı bir platform olduğu için içerik üreticileri ve yayıncılar için de cazip bir ortam sunmaktadır. Bu sayede, yayıncılar, içeriklerini geniş bir kullanıcı kitlesine eriştirebilirler.

Facebook, ayrıca bir sosyal paylaşım platformudur. Kullanıcılar, fikirlerini, duygularını ve anılarını paylaşabilirler. Ayrıca, kullanıcılar arasında etkileşimli bir platformdur. Kullanıcılar, paylaşımlara "beğenme", yorum yapma veya paylaşma gibi şekillerde etkileşimde bulunabilirler. Facebook, günümüzde dünya çapında en yaygın kullanılan sosyal medya platformlarından biridir. Bu platform kullanıcılar için sosyal iletişim, haberler ve bilgi paylaşımı, eğlence ve işletmeler için pazarlama ve reklam amaçları için kullanılabilir.

3.4 Sosyal Medya Neden Önemlidir

Sosyal medya, insanlar arasında iletişimi ve bağlantı kurmayı kolaylaştıran bir araçtır. Bu nedenle, sosyal medya kullanıcıların iş ve kişisel ilişkilerini geliştirmelerine, haberleşmelerine ve bilgiye erişmelerine olanak tanır. Ayrıca, sosyal medya, işletmelerin müşterileriyle iletişim kurmasına ve pazarlama faaliyetlerini gerçekleştirmesine olanak tanır.

Sosyal medya ayrıca, toplumda oluşan güncel olaylar ve haberler hakkında bilgi sahibi olmak için önemli bir kaynaktır. Ayrıca, sosyal medya kullanıcıların kültür, sanat, müzik ve diğer ilgi alanları hakkında bilgi edinmelerine ve diğer kişilerle

paylaşmalarına olanak tanır. Özellikle, sosyal medya kullanıcıların yerel ve uluslararası topluluklarla iletişim kurmalarına ve bu topluluklarla birlikte çalışmalarına olanak tanır.

Son olarak, sosyal medya ayrıca, kişisel ve profesyonel olarak kendinizi tanıtmak için önemli bir araçtır. Kişisel bir blog, Instagram hesabı veya LinkedIn profili oluşturarak, insanların sizi ve işinizi tanıyabilecekleri bir platform oluşturabilirsiniz. Sosyal medya ayrıca, müşteri hizmetleri ve müşteri memnuniyeti konusunda da önemlidir. İşletmeler, sosyal medya kanallarını kullanarak müşterilerine anında cevap verebilir ve müşteri şikâyetlerini çözüme kavuşturabilirler. Ayrıca, sosyal medya kanalları, işletmeler için geri bildirim alma ve ürün ve hizmetleri geliştirme konusunda önemli bir kaynaktır.

Son olarak, sosyal medya ayrıca, e-ticaret ve dijital pazarlama konusunda önemli bir araçtır. İşletmeler, sosyal medya kanallarını kullanarak ürünlerini ve hizmetlerini tanıtmak, müşterilerle etkileşim kurmak ve satış yapmak için bir platform sağlar. Sosyal medya ayrıca, işletmelerin reklam ve pazarlama kampanyalarını hedefleyebilecekleri bir kitleye erişmelerine olanak tanır.

3.5 Sosyal Medyada En Çok Kullanılan Platform Olan Twitter'ın Yaptığımız Araştırma Açısından Literatüre Katkısı

Twitter, kullanıcıların kısa mesajlar olarak bilinen "tweet"leri paylaşabildiği bir sosyal medya platformudur. Bu nedenle, Twitter özellikle haber, güncel olaylar ve güncel konular hakkında bilgi sahibi olmak için önemli bir kaynaktır. Ayrıca, Twitter, kullanıcıların duygularını, görüşlerini ve fikirlerini paylaşmalarına olanak tanır. Twitter'da hakaret içeren cümleler, dil ve duygu analizi araştırmaları için önemli bir veri kaynağıdır. Bu cümleler, dil ve duygu analizi araştırmaları için önemli veri kümelerini oluşturur ve araştırmacıların dil ve duygu analizi yöntemlerini geliştirmelerine yardımcı olur. Bu tür cümlelerin analizi, araştırmacıların insanların duygusal durumlarını, tutumlarını ve görüşlerini belirlemelerine yardımcı olur. Bu cümleler ayrıca, insanların nasıl bir dil kullandıklarını ve hangi duyguları ifade ettiklerini analiz etmek için kullanılabilir. Bu veriler, sosyal medya kullanımının etkilerini ve insanların duygusal yanıtlarını anlamak için önemli bilgiler sağlar.

Sonuç olarak, Twitter hakaret içeren cümleler ve duygu analizi, literatüre dil ve duygu analizi alanında önemli katkılar sağlamaktadır. Bu katkılar, insanların duygusal durumlarını, tutumlarını ve görüşlerini belirlemek için kullanılabilecek yöntemleri

geliştirmekte ve sosyal medya kullanımının etkilerini anlamak için önemli bilgiler sağlamaktadır.

Twitter hakaret içeren cümleler ve duygu analizi literatürü ayrıca, çeşitli toplumsal ve sosyal sorunların önlenmesine katkıda bulunabilir. Örneğin, cinsiyetçi, ırkçı veya homofobik dil kullanımının yaygınlığını ve etkilerini belirlemek ve bunların önlenmesine yönelik çalışmalar yapmak için kullanılabilir. Aynı zamanda, sosyal medya platformlarındaki hakaret içeren cümleler, insanların sağlıklı bir şekilde iletişim kurmasının engellenmesine neden olabilir, bu nedenle önlenmesi önemlidir. Twitter hakaret içeren cümleler ve duygu analizi literatürü ayrıca, dil ve duygu analizi alanında yapılan araştırmaların sonuçlarının gerçek hayatta uygulanmasına olanak tanır. Örneğin, dil ve duygu analizi yöntemleri kullanılarak tasarlanan bir sistem, sosyal medya platformlarında hakaret içeren cümleleri otomatik olarak tespit edebilir ve bunları önleyebilir. Ayrıca, dil ve duygu analizi yöntemleri, sosyal medya platformlarındaki kullanıcıların duygusal durumlarını ve tutumlarını takip etmek ve bu bilgileri kullanmak için kullanılabilir. Sonuç olarak, Twitter hakaret içeren cümleler ve duygu analizi literatürü, dil ve duygu analizi alanında önemli katkılar sağlamaktadır. Bu katkılar, toplumsal ve sosyal sorunların önlenmesine, gerçek hayatta uygulanabilecek sistemlerin tasarlanmasına ve insanların duygusal durumlarını ve tutumlarını takip etmek için kullanılabilecek yöntemleri geliştirmektedir.

3.6 Sosyal Medya Analitiği Üzerinden Verinin Önemi

Sosyal medyada analitik, sosyal medya platformlarındaki verileri toplamak, analiz etmek ve anlamak için kullanılan bir yöntemdir. Sosyal medya analitik veriler, işletmeler, markalar, ajanslar ve diğer kurumlar için önemli bir bilgi kaynağıdır.

Sosyal medya analitik veriler, işletmelerin sosyal medya performanslarını ve etkilerini ölçmelerine, müşteri davranışlarını ve trendlerini analiz etmelerine ve pazarlama ve satış stratejilerini geliştirmelerine olanak tanır. Örneğin, sosyal medya analitik veriler, işletmelerin hangi ürünlerin veya hizmetlerin en popüler olduğunu, hangi takipçilerin en etkili olduğunu ve hangi zaman dilimlerinde en yüksek etkileşim oranlarının olduğunu belirlemelerine olanak tanır. Sosyal medya analitik veriler ayrıca, işletmelerin müşteri memnuniyetini ve sadakatini ölçmelerine ve müşteri ilişkilerini geliştirmelerine

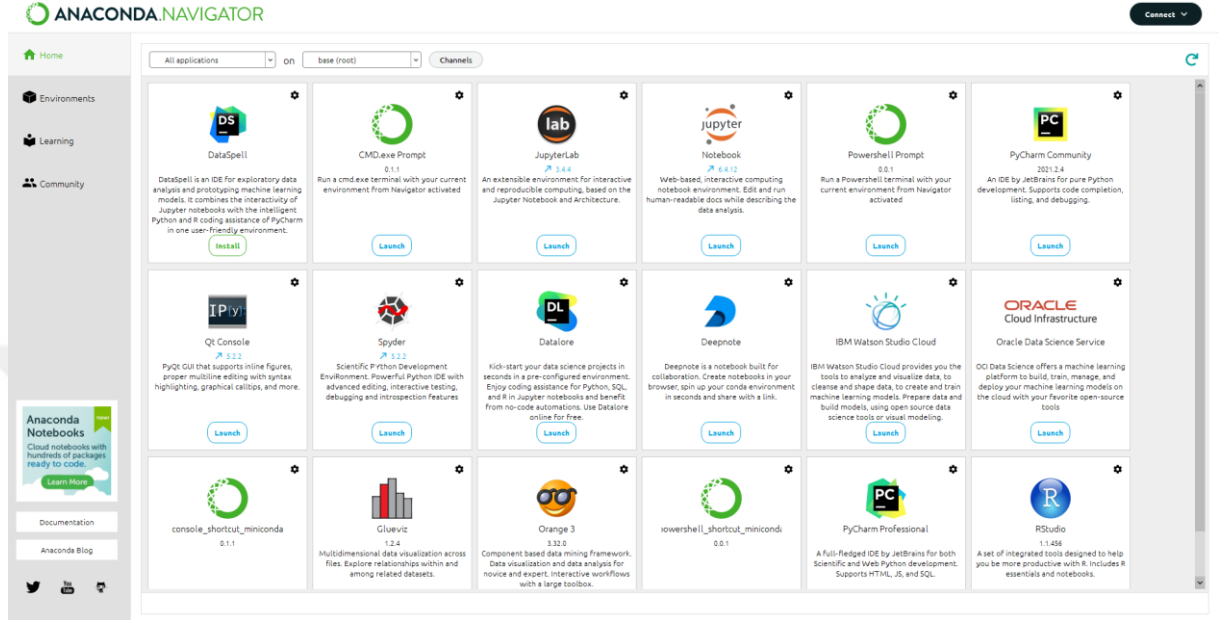
yardımcı olur. Örneğin, sosyal medya analitik veriler, işletmelerin müşteri şikâyetlerini ve önerilerini takip etmelerine ve bunları çözüme kavuşturmalarına olanak tanır.

Sonuç olarak, sosyal medya analitik veriler, işletmeler, markalar ve diğer kurumlar için önemli bir bilgi kaynağıdır. Sosyal medya analitik veriler, işletmelerin sosyal medya performanslarını ve etkilerini ölçmelerine, müşteri davranışlarını ve trendlerini analiz etmelerine ve pazarlama ve satış stratejilerini geliştirmelerine olanak tanır. Bu veriler ayrıca, işletmelerin müşteri memnuniyetini ve sadakatini ölçmelerine ve müşteri ilişkilerini geliştirmelerine yardımcı olur. Sosyal medya analitik veriler, ayrıca, işletmelerin reklam ve pazarlama kampanyalarının etkililiğini ölçmelerine ve optimize etmelerine olanak tanır. Örneğin, sosyal medya analitik veriler, işletmelerin hangi reklam veya pazarlama kampanyalarının en yüksek tıklama oranlarına veya satışlara neden olduğunu belirlemelerine olanak tanır. Bu veriler ayrıca, işletmelerin hangi demografik gruplara yönelik reklam ve pazarlama kampanyalarının daha etkili olduğunu veya hangi sosyal medya kanallarının daha etkili olduğunu belirlemelerine olanak tanır. Sosyal medya analitik veriler ayrıca, işletmelerin rakiplerinin performanslarını ve etkilerini analiz etmelerine olanak tanır. Örneğin, sosyal medya analitik veriler, işletmelerin rakiplerinin hangi ürünlerin veya hizmetlerin en popüler olduğunu, hangi takipçilerin en etkili olduğunu ve hangi zaman dilimlerinde en yüksek etkileşim oranlarının olduğunu belirlemelerine olanak tanır. Bu veriler ayrıca, işletmelerin rakiplerinin hangi reklam ve pazarlama kampanyalarının daha etkili olduğunu veya hangi sosyal medya kanallarının daha etkili olduğunu belirlemelerine olanak tanır. Sonuç olarak, sosyal medya analitik veriler, işletmeler, markalar ve diğer kurumlar için önemli bir bilgi kaynağıdır. Sosyal medya analitik veriler, işletmelerin sosyal medya performanslarını, etkilerini, reklam ve pazarlama kampanyalarının etkililiğini, rakiplerin performanslarını ve etkilerini analiz etmelerine ve optimize etmelerine olanak tanır. Bu veriler ayrıca, işletmelerin müşteri memnuniyetini, sadakatini, reklam ve pazarlama kampanyalarının etkililiğini, rakiplerin performanslarını ve etkilerini ölçmelerine ve müşteri ilişkilerini geliştirmelerine yardımcı olur

BÖLÜM 4: BULGULAR

4.1 Metotların ve Materyallerin Kullanılması

Anaconda navigator(Anaconda3) üzerinden spyder 5.2.2 IDE'si üzerinden geliştirmeler yapıldı.



Şekil 15: Geliştirme ortamında kullanılan Anaconda navigator



Şekil 16: Uygulama aşamaları diyagramı

4.2 Geliştirmede Kullandığımız Gerekli Kütüphaneler

```
#warnings
import warnings
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import time
import re
import nltk
nltk.download('wordnet')
nltk.download('omw-1.4')
from unicode_tr import unicode_tr
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from nltk.stem import WordNetLemmatizer
#from sklearn.train_test_split import train and fit models
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.feature_extraction.text import TfidfVectorizer
#model selection
from sklearn.metrics import confusion_matrix, accuracy_score
#word plot
from wordcloud import WordCloud
```

4.3 Geliştirmede Kullandığımız Gerekli Kütüphanelerin Açıklamaları

“import warnings”

Warning/Uyarıları kapatmak için kullandığımız kütüphanedir.

“import pandas as pd”

Data Frame ile ilgili analizler yapabilmemiz için gereken kütüphanedir.

“import numpy as np”

Vektörleri içerisinde tutuyor. Matematiksel işlemleri kullandığımız kütüphanedir.

“import seaborn as sns”

Görselleştirme için kullandığımız kütüphanedir.

“import matplotlib.pyplot as plt”

Görselleştirme için kullandığımız kütüphanedir.

“import time”

Zaman ve tarihle ilgili işlemler yapmak için kullanılan fonksiyonlar içerir

“import re”

Düzenli ifadeler (regular expressions) kullanarak metin içinde arama ve düzenleme işlemlerini gerçekleştirmek için kullanılır.

“import nltk” NLTK (Natural Language Toolkit) dil işleme işlemleri için kullanılan bir kütüphanedir. Bu kütüphane, dil işleme işlemleri için birçok fonksiyon ve araçlar içerir. Örneğin, metinleri tokenize etmek, POS (part-of-speech) etiketleme, named entity recognition, sentiment analysis, text summarization gibi işlemleri gerçekleştirmek için kullanılabilir.

“from sklearn.preprocessing import StandardScaler”

Preposess yapıyoruz. Standardizasyon yapabilmemiz için bu kütüphaneyi kullanmamız gereklidir.

“from sklearn.model_selection import train_test_split”

Train test split dediğimiz ise train yaparak eğitiriz, test alanında ise test yaparız ve bunları ayırırız. Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir fonksiyonudur. Eğitim verileri, makine öğrenimi modelinin öğrenmesi için kullanılırken, test verileri ise modelin performansının değerlendirilmesi için kullanılır. Bu veri bölme işlemi, modelin gerçek hayatta nasıl performans göstereceğini daha doğru bir şekilde tahmin etmek için yapılır.

“from sklearn.metrics import confusion_matrix, accuracy_score”

Modelimizi eğittikten sonra çıkan sonuçları değerlendirebilmek için metrik kullanmamız gerekir. En temel metriklerden accuracy’i kullanacağız. Yani örneğin %30 hakaret içerikli, %70 saldırgan gibi örnek verebiliriz. Diğer ise confusion matrix’tir. Kullanmamızın sebebi ise accuracy değerini kullanırken değerlendirme yapabiliyoruz. Değerlendirmenin nedenini bulmak için confusion matrix’ini kullanmamız gerekir.

“from matplotlib.colors import ListedColormap”

Çıkan sonuçlarımızı görselleştirmemiz için kullanacağız.

“from unicode_tr import unicode_tr”

Türkçe metin işlersek eğer bu kütüphaneyi de eklememiz gerekir. Türkçe metin ya da veri işlemek için fonksiyonlar, sınıflar içerebilir.

“from nltk.corpus import stopwords”

NLTK (Natural Language Toolkit) adlı Python dilinde kullanılan bir dil işleme kütüphanesinde yer alan, dilden dille değişebilen "durak kelimeler" (stop words) içeren bir veri kümesidir. Bu kütüphane, metinlerdeki bu kelimeleri belirli bir amaç için filtrelemek için kullanılabilir. Örneğin, bir metin analitik uygulamasında, durak kelimelerin metinden çıkarılması, geri kalan kelimeler arasındaki anlam ilişkilerini daha

iyi görüntülemek için yararlı olabilir. Engellenecek sözcükleri, bir dilde yaygın olarak kullanılan metin sınıflandırma veya bilgi çıkarma gibi doğal dil işleme görevleri sırasında tipik olarak metinden kaldırılan sözcüklerin bir listesini içerir. Bu kelimeler, herhangi bir faydalı bilgi içermedikleri ve genellikle doğal dil işleme algoritmalarının performansını artırmak için kullanıldığı için "durdurma kelimeleri" olarak kabul edilir.

“from sklearn.feature_extraction.text import CountVectorizer”

Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, metinleri vektörlere dönüştürmek için kullanılır. Bu işlem, metinleri sıklık bazında sayılar vektörlerine dönüştürür. Bu sayılar, metinde belirli bir kelime ya da terimin kaç kez geçtiğini gösterir. Bu vektörler, metinleri sınıflandırmak, kategorize etmek veya diğer makine öğrenimi algoritmalarında kullanmak için kullanılabilir. Özellikle metin sınıflandırma, metin kümeleme gibi işlemler için kullanılabilir.

“from nltk.stem import WordNetLemmatizer”

NLTK (Natural Language Toolkit) adlı Python dilinde kullanılan bir dil işleme kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, metinlerdeki kelimeleri köklerine dönüştürmek için kullanılır. Bu işlem, kelimelerin çekim ekleri, fiil şekilleri gibi biçimlerinden arındırılmasını sağlar ve böylece metinlerdeki anlamsal bilgi daha iyi çıkarılabilir. Bu kökler, metinleri sınıflandırmak, kategorize etmek veya diğer dil işleme algoritmalarında kullanmak için kullanılabilir. Özellikle metin madenciliği ve metin analitik gibi işlemler için kullanılabilir.

“from sklearn.ensemble import RandomForestClassifier”

Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, rastgele orman sınıflandırıcısı (Random Forest Classifier) algoritmasını uygulamak için kullanılır. Rastgele orman, birden fazla karar ağacının bir araya getirilmesiyle oluşan bir ensemble yöntemidir. Her bir karar ağacı, veri kümesinden rastgele seçilmiş alt veri kümelerini kullanarak eğitilir ve sonuç olarak, tahminlerin ortalaması alınır. Bu yöntem, veri kümesinde çok sayıda değişken olsa bile, iyi performans gösterebilir ve overfitting (aşırı uyum) problemine karşı dayanıklıdır.

“from sklearn.linear_model import LogisticRegression”

Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, lojistik regresyon algoritmasını uygulamak için kullanılır. Lojistik regresyon, sınıflandırma problemlerinde kullanılan bir algoritmadır. Bu algoritma, veri kümesindeki değişkenler arasındaki ilişkileri inceler ve veri noktalarının hangi sınıfa ait olduğunu tahmin etmek için bir sınır çizgisi oluşturur. Lojistik regresyon, özellikle iki sınıflı sınıflandırma problemlerinde yaygın olarak kullanılır ve özellikle değişkenler arasındaki ilişkiler lineer olduğunda iyi performans gösterir.

“from sklearn.neighbors import KNeighborsClassifier”

Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, K en yakın komşu (K-nearest neighbors, KNN) sınıflandırma algoritmasını uygulamak için kullanılır. KNN algoritması, veri kümesindeki veri noktalarının sınıflandırılması için kullanılır. Bu algoritma, yeni gelen bir veri noktasının sınıfını tahmin etmek için, veri kümesindeki en yakın k veri noktasının sınıfını inceler. Bu algoritma, veri kümesinde çok sayıda değişken olsa bile, iyi performans gösterebilir ve özellikle veri noktaları arasındaki ilişkiler non-lineer olduğunda iyi sonuçlar verir.

“from sklearn.tree import DecisionTreeClassifier”

Scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, karar ağacı algoritmasını uygulamak için kullanılır. Karar ağacı algoritması, sınıflandırma ve regresyon problemlerinde kullanılan bir algoritmadır. Bu algoritma, veri kümesindeki değişkenler arasındaki ilişkileri inceler ve veri noktalarının hangi sınıfa ait olduğunu tahmin etmek için bir ağaç yapısı oluşturur. Her bir karar noktasında, ağaç yapısındaki bir düğümde, bir değişken üzerinde bir karar verilir ve veri noktaları o değişken üzerinde belirli bir değere sahip olduğunda ağaç yapısının hangi yolunu izleyeceği belirlenir. Bu algoritma, veri kümesinde çok sayıda değişken olsa bile, iyi performans gösterebilir ve özellikle veri noktaları arasındaki ilişkiler non-lineer olduğunda iyi sonuçlar verir.

“from sklearn.naive_bayes import GaussianNB”

scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, Gaussian Naive Bayes algoritmasını uygulamak için kullanılır. Gaussian Naive Bayes algoritması, sınıflandırma problemlerinde kullanılan bir algoritmadır. Bu algoritma, veri kümesindeki değişkenler arasındaki ilişkileri inceler ve veri noktalarının hangi sınıfa ait olduğunu tahmin etmek için bir sınıflandırıcı oluşturur. Algoritma, her değişken için verilerin normal dağılımını varsayar ve veri noktalarının her bir değişken için ait olduğu normal dağılımın parametrelerini tahmin eder. Bu algoritma, özellikle veri noktaları arasındaki ilişkiler lineer olduğunda iyi performans gösterir.

“from sklearn.feature_extraction.text import TfidfVectorizer”

scikit-learn adlı Python dilinde kullanılan bir makine öğrenimi kütüphanesinde yer alan bir sınıfı temsil eder. Bu sınıf, metin verilerini sayısal bir vektör formuna dönüştürmek için kullanılan TF-IDF (Term Frequency-Inverse Document Frequency) algoritmasını uygulamak için kullanılır. Bu algoritma, metin verilerini özellik vektörü olarak temsil etmek için kullanılır. Bu vektörler, metin verilerinde yer alan kelimelere göre oluşturulur ve her bir kelime için metin verilerinde o kelimenin görülme sıklığını (Term Frequency) ve metin verileri içinde o kelimenin nadir görülme oranını (Inverse Document Frequency) temsil eder. Bu algoritma, metin verileri için özellik seçimi yapmak için kullanılır ve metin verilerinde anlamlı kelimelerin belirlenmesini sağlar.

“from wordcloud import WordCloud”

Bu kütüphane, metin verilerinde yer alan kelimelerin görsel bir şekilde sunulmasını sağlar. Kullanıcılar, metin verilerinde yer alan kelimelerin büyüklüklerini ve renklerini değiştirerek, metin verilerinde hangi kelimelerin daha fazla yer aldığını görsel olarak anlayabilir. Bu kütüphane, metin verilerinin analizinde ve özellik seçiminde kullanılan bir araçtır. Özellikle, "WordCloud" sınıfı, metin verilerinde yer alan kelimelere göre oluşturulan kelime bulutlarını oluşturmak için kullanılır. Bu kelime bulutları, metin verilerinde yer alan kelimelerin büyüklüklerini ve renklerini kullanarak görsel olarak sunulur. Geliştirme ortamlarında genellikle bu kütüphane tercih edilir. Veri setinde görselleştirme ve renklendirmesi ile kullanıcıya kolaylık sağlar.

4.4 Geliştirmede Kullanılan Kodların Açıklamaları

Veri setini hakaret içerikli cümlelerin, saldırgan cümlelerin ve duygu durumunun pozitif olduğu yani ne hakaret içerikli ne de saldırgan bir tavır olmadığı cümlelerin olduğu veri seti ile yaptım. Veri setinin kaggle veri set platformunu kullanarak çalışma yapıldı. Veri setindeki cümlelere göre üç değer üzerinde tahminde bulunduk. Bu üç değer:

- **Hate:** Hakaret içeren cümleleri, bu değer üzerinden makine öğrenme tekniklerine göre işleyip geliştirilen uygulama üzerinden dönüş yapacak.
- **Offensive:** Saldırgan olan cümleleri, bu değer üzerinden makine öğrenme tekniklerine göre işleyip geliştirilen uygulama üzerinden dönüş yapacak.
- **None:** Nötr olan cümleleri pozitif olarak cümleleri ise bu değer üzerinden makine öğrenme tekniklerine göre işleyip geliştirilen uygulama üzerinden dönüş yapacak.

```
Dic={0:'Hate',  
      1:'Offensive',  
      2:'None'}
```

Key ▲	Type	Size	Value
0	str	4	Hate
1	str	9	Offensive
2	str	4	None

Şekil 17: Değer alanları

Tablo 2

Veri seti alan bilgileri

RangeIndex: 24783 entries, 0 to 24782

Data columns (total 7 columns):

#	Column	Non-Null	Count	Dtype
0	Unnamed: 0	24783	non-null	int64
1	Count	24783	non-null	int64
2	hate_speech	24783	non-null	int64
3	offensive_language	24783	non-null	int64
4	Neither	24783	non-null	int64
5	Class	24783	non-null	int64
6	Tweet	24783	non-null	Object

dtypes: int64(6), object(1)

memory usage: 1.3+ MB

Veri seti toplamda 7 kolondan 24783 kayıttan oluşmaktadır. Veri setinde bizim için önemli alanlar ise index, class ve tweet alanlarıdır.

Bu veri setinden bir tane drop etmek istediğimiz feature var. Unnamed: 0 column name.

```
data.drop(['Unnamed: 0'],inplace=True,axis=1)
```

Unnamed: 0 isimli kolonu drop ettik. İnplace kendi ekseninde, axis=1 dediğimizde kolonu drop etmiş oluruz.

Class alanını Target olarak değiştirerek, etiketleştirmiş oluruz.

index	count	hate_speech	offensive_language	neither	target	tweet
0	3	0	0	3	2	!!! RT @mayssolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
1	3	0	3	0	1	!!!! RT @leew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!! RT @urkindoffrend Dawg!!!! RT @80baby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!! RT @C_GAnderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!! RT @shenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &457361;
5	3	1	2	0	1	!!!!!! RT @Madison_x: The shit just blows me...clade you so faithful and down for somebody but still fucking with hoes! &4128514;&4128514;"
6	3	0	3	0	1	!!!!!! @_BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
7	3	0	3	0	1	!!!!&46220@Selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!&46221;
8	3	0	3	0	1	" & you might not get ya bitch back & that's that "
9	3	1	2	0	1	@tychmaka_ "housdco atcloud: tagging notson
10	3	0	3	0	1	" Weeks is a bitch she curves everyone " lol I walked into a conversation like this. Sm

Şekil 18: Veri setinin ilk 10 satırı

Tablo 3

Veri setindeki mevcut durum üzerinden bilgileri

RangeIndex: 24783 entries, 0 to 24782

Data columns (total 6 columns):

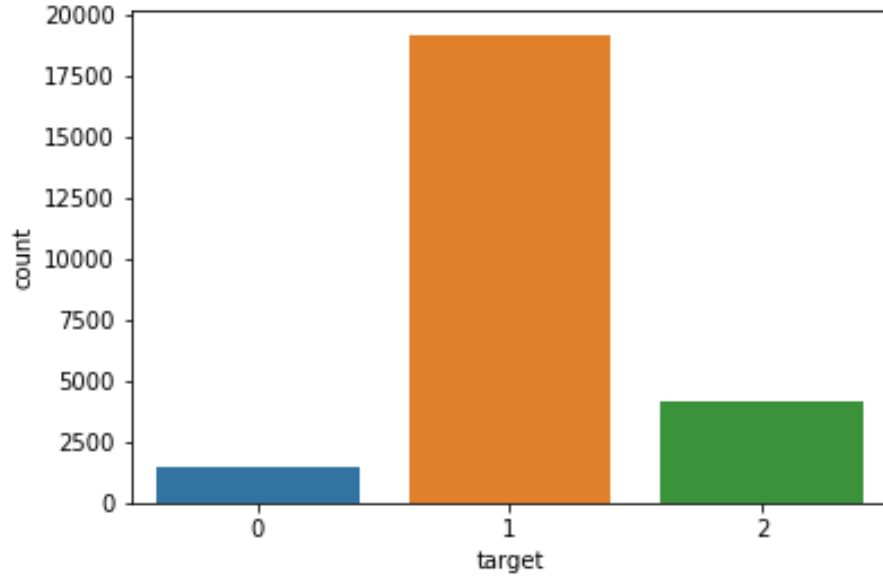
#	Column	Non-Null	Count	Dtype
1	Count	24783	non-null	int64
2	hate_speech	24783	non-null	int64
3	offensive_language	24783	non-null	int64
4	Neither	24783	non-null	int64
5	Class	24783	non-null	int64
6	Tweet	24783	non-null	object

dtypes: int64(5), object(1)

memory usage: 1.1+ MB

data.info()

Kullanılan kolonlar hakkında bilgiyi çekeriz. Missing value değerleri varsa görülebilir. Toplamdaki değer 24783 geldiğinden dolayı herhangi bir missing value yoktur. Çıkan tabloya göre toplamda beş kolon numerik değerlerden, bir kolon ise string değerlerden oluşmuştur.



Şekil 19: Veri setinde target kolonunun söylemlere göre görselleştirilmiş dağılımı

```
sns.countplot(data["target"])
```

Snsborn kütüphanesinden countplot methoda göre data frame'nin alanlarını görmüş oluruz. Grafiği incelediğimizde yaklaşık değerleri görürüz. Bu yüzden aşağıdaki yöntem ile net sayıları çıkarmış oluruz.

Index	target
1	19190
2	4163
0	1430

Şekil 20: Veri setinde target kolonunun söylemlere göre numerik dağılımı

```
netsayilar=data.target.value_counts()
```

19190 tane 1, saldırgan içerikli cümle,

4163 tane 2, none/possitive cümle,

1430 tane 0, hakaret içerikli cümle olduğunu tespit etmiş oluruz.

describe - DataFrame

Index	count	hate speech	offensive language	neither	target
count	24783	24783	24783	24783	24783
mean	3.24347	0.280515	2.41371	0.549247	1.11028
std	0.88306	0.631851	1.39946	1.1133	0.462089
min	3	0	0	0	0
25%	3	0	2	0	1
50%	3	0	3	0	1
75%	3	0	3	0	1
max	9	7	9	9	2

Şekil 21: Veri setindeki alanların total duruma göre görülmesi

```
describe = data.describe()
```

İlk sırada yer alan count, veri setinde feature'lerin içerisinde bulunan 24783 sample sayıdır.

Mean dediğimiz ise ortalamalarıdır. Target alanındaki ortalamamız 1.11028'dir.

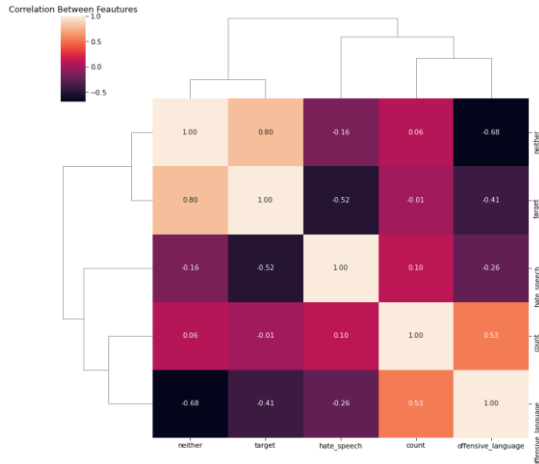
corr_matrix - DataFrame

Index	count	hate_speech	offensive_language	neither	target
count	1	0.101926	0.534604	0.0633261	-0.0100286
hate_speech	0.101926	1	-0.261808	-0.157599	-0.515732
offensive_language	0.534604	-0.261808	1	-0.684406	-0.41044
neither	0.0633261	-0.157599	-0.684406	1	0.800687
target	-0.0100286	-0.515732	-0.41044	0.800687	1

Şekil 22: Veri setinin korelasyon matrisine göre değerleri

```
corr_matrix = data.corr()
```

Korelasyon matrisini bu şekilde anlamamız zor bu yüzden görselleştirmemiz gereklidir.



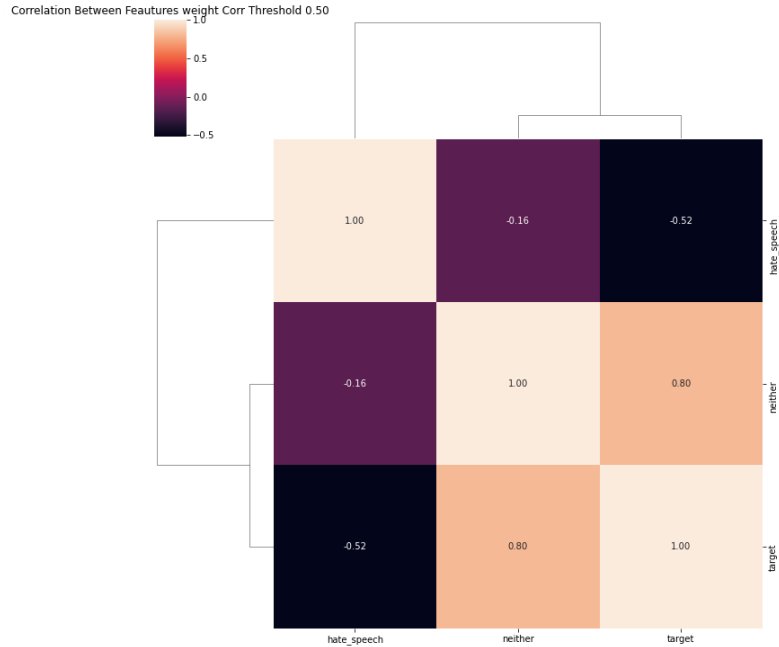
Şekil 23: Veri setindeki alanlara göre korelasyon matrisinin görselleştirilmiş hali

```
corr_matrix = data.corr()  
sns.clustermap(corr_matrix,annot=True, fmt=".2f")
```

Sns kütüphanesinden clustermap methodu kullanarak korelasyon matrisinin içerisinde ki sayılar. Bu sayıları görürken virgülden sonra iki değer görmek istiyoruz.

Korelasyona örnek vermek gerekirse: Örneğin iki adet feature var. Bu feature'lar arası ilişkiye bakarız. Eğer bu iki feature 1 ise %100 doğru orantılıdır. Eğer bu iki feature -1 ise %100 ters orantılıdır. Eğer ki 0 ise bir ilişki yoktur diyebiliriz.

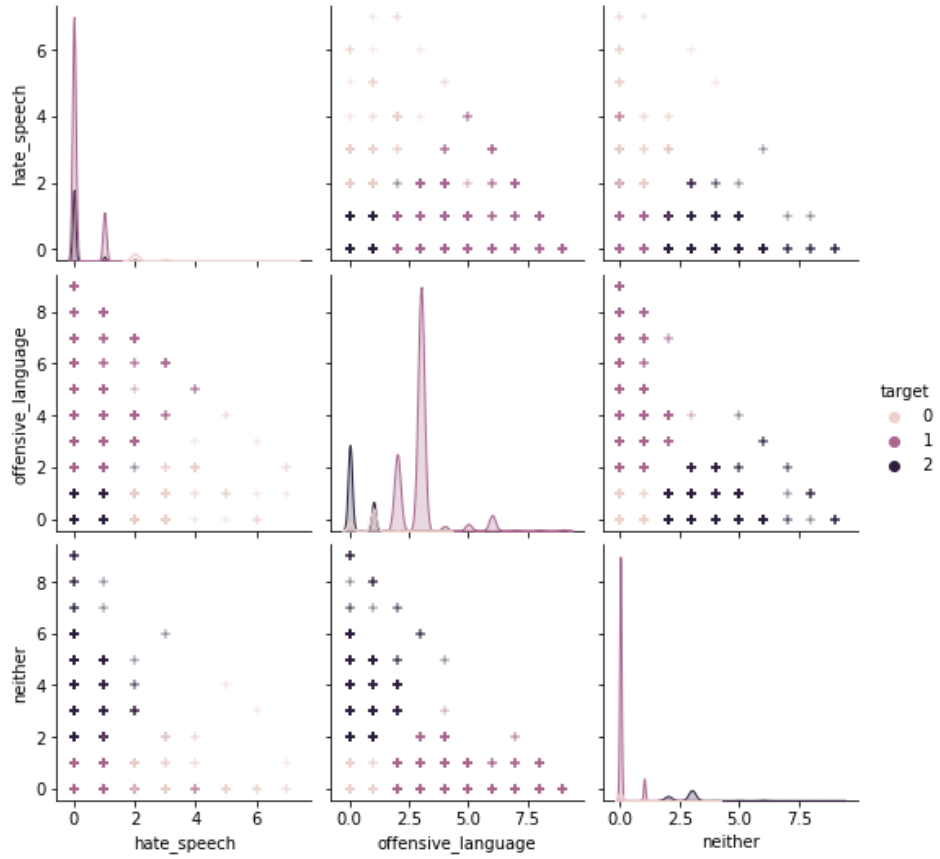
Grafiği incelersek ağaç yapısının iki dalından aşağıya doğru indiğimizde ki denk gelen kolon neither ve target bu iki feature birbirleriyle ilişkilidir. 0.80 değerini aldığını da görmekteyiz. Buradan çıkarabilecek sonuç ise; birbirleriyle alakalıysalar modele olan katkı aynıdır. 0.80 değerli neither feature'sini makine öğrenim tekniklerine göre eğitirsek çıkan katkı aynı olacak, değişme olmayacaktır. Bu yüzden makine öğrenim tekniklerinden çeşitliliğe gitmek gereklidir. Örneğin count ile hate speech arasında %10'luk bir ilişki var bu yüzden birbirleri arasında korelasyon çok fazla yoktur.



Şekil 24: Threshold değeri girilerek korelasyon matrisinin görselleştirilmiş hali

```
threshold = 0.50
filtre = np.abs(corr_matrix["target"]) > threshold
corr_features = corr_matrix.columns[filtre].tolist()
sns.clustermap(data[corr_features].corr(),annot=True,fmt=".2f")
plt.title("Correlation Between Features weight Corr Threshold 0.50")
```

Threshold değeri girerekte 1'e yakın olan ilişkili features'ler getirilebilmektedir.



Şekil 25: Veri setindeki features'lerin histogram olarak görünümü

```
sns.pairplot(data[corr_features],diag_kind="kde",markers="+",hue="target")
plt.show()
```

Data'nın içerisinde sadece corr_features'lere bakarız. Histogram olarak yaparız. Marker + yaparız. Sınıflandırmamız ise Target seçerek görselleştirme işlemini yapılabilir. %50

threshold değerine göre grafik oluşmuştur. Target ve diğer corr_features arasındaki ilişkileri görebiliriz.

dt_transformed - DataFrame

Index	target	tweet
0	2	!!! RT @mayasolovely: As a woman you should...
1	1	!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	1	!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sb...
3	1	!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	1	!!!!!! RT @ShenikaRoberts: The shit ...
5	1	!!!!!!"@T_Madison_x: The shit j...
6	1	!!!!!!"@__BrighterDays: I can not just sit ...
7	1	!!!!“@selfiequeenbri: cause I'm tired...
8	1	" & you might not get ya bitch back & thats that "
9	1	@nylmmxxx_ :hobbies include: fighting nd...
10	1	bitch " Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh

Şekil 26: dt_transformed tablosunun ilk 10 satırı

```
dt_transformed = data[['target','tweet']]
y = dt_transformed['target'].values
print(y)
print(dt_transformed)
```

Target ve Tweet kolonlarını kullanıp, bu iki kolonların değerlerini alır. Sonrasında transform işlerini yapar.

```
ax = dt_transformed.groupby('target').count().plot(kind='bar', title='Distribution of data',
legend=False)
```

Bu kod, 'target' değişkenine göre gruplama yaparak verilerin dağılımını görselleştirir. Distribution data başlığı altında, verilerin dağılımı için bar grafiği kullanılır ve ax değişkeninin içine atar.

```
ax.set_xticklabels(['hate','offensive','None'],rotation=0)
```

Bu kod eksenin x noktalarının etiketlerini 'Hate', 'Offensive', 'None' şeklinde ayarlar ve etiketlerin dönüş açısı 0 derece olacak şekilde ayarlanır.

```
def preprocess(textdata):  
    processedText = []
```

Bu fonksiyon, girdi olarak verilen metin verilerini işler ve işlenmiş metin verilerini saklamak için bir liste oluşturur.

```
wordLemm = WordNetLemmatizer()
```

Bu satır, kelimeleme işlemleri için WordNetLemmatizer nesnesini tanımlar.

```
urlPattern      = r"((http://)[^ ]*|(https://)[^ ]*|( www\.)[^ ]*)"
userPattern     = '@[^\s]+'
alphaPattern    = "[^a-zA-Z0-9]"
sequencePattern = r"(\1\1+"
seqReplacePattern = r"\1\1"
newlinepattern  = '\n'
```

Bu satırlar, metnin işlenmesi sırasında kullanılacak regex (düzenli ifade) desenlerini tanımlar:

urlPattern: http:// veya https:// ile başlayan veya www. ile başlayan URL adreslerini belirler.

userPattern: @ işareti ile başlayan kullanıcı adlarını belirler.

alphaPattern: Alfanümerik karakterler dışındaki karakterleri belirler.

sequencePattern: Üç veya daha fazla tekrar eden karakter dizilerini belirler.

seqReplacePattern: Üç veya daha fazla tekrar eden karakter dizilerinin iki tekrar eden karakter ile değiştirilmesi için kullanılır.

newlinepattern: Metnin içindeki yeni satır karakterlerini belirler.

```
for text in textdata:  
    text = text.lower()
```

for döngüsü, verilen metin verilerinin hepsini döner ve her bir metnin karakterlerini küçük harfe çevirir.

```
#  
text = re.sub(urlPattern, 'URL', text)
```

Veri setindeki tüm URL'leri 'URL' şekline değiştiriyor

```
# Replace @USERNAME to 'USER'.  
text = re.sub(userPattern, 'USER', text)  
# Replace all non alphabets.  
text = re.sub(alphaPattern, " ", text)  
# Replace all new line with " "  
text = re.sub(newlinepattern, " ", text)  
# Replace 3 or more consecutive letters by 2 letter.  
text = re.sub(sequencePattern, seqReplacePattern, text)
```

İlk satır, "@USERNAME" olarak belirtilen tüm text'leri "USER" şeklinde değiştirir.

İkinci satır, alfabetik olmayan tüm karakterleri boşluklarla değiştirir.

Üçüncü satır, tüm yeni satır karakterlerini boşluklarla değiştirir.

Dördüncü satır, üç veya daha fazla ardışık harf şeklindeki dizileri iki harfle değiştirir.

```
textwords = ""
for word in text.split():
    # Checking if the word is a stopword.
    #if word not in stopwordlist:
        if len(word)>1:
            # Lemmatizing the word.
            word = wordLemm.lemmatize(word)
            textwords += (word+' ')

processedText.append(textwords)

return processedText
```

Bu, bir metin dizisi için verilen işlemleri yapar. Her kelime için, durdurma kelimesi olup olmadığı kontrol edilir. Eğer durdurma kelimesi değilse, kelime uzunluğu 1'den büyükse, kelime lemmatize edilir ve işlenmiş kelime dizisine eklenir. İşlenmiş metin dizisi döndürülür.

```
t = time.time()
corpus = preprocess(dt_transformed['tweet'])
print(f'Text preprocessing complete.')
print(f'Time taken: {round (time.time()-t)} second s')
```

Bu kod, 'tweet' adlı pandas veri çerçevesi sütunundaki bir metin korpusunu ön işleme yapar. Preprocess fonksiyonu, metin verilerinin daha sonraki analize hazırlanması için temizlenmesini ve hazırlanmasını sağlar. İşlemenin zamanı zaman modülü kullanılarak hesaplanır ve sonunda yazdırılır. "Yuvarlama" fonksiyonu, alınan zamanı en yakın tam sayıya yuvarlar.

Çıktısı:

```
[24783 rows x 2 columns]
Text preprocessing complete.
```

Time taken: 2 second s

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state = 0)
```

X_train, X_test, y_train, y_test adlı 4 değişkenin oluşturulduğu train_test_split fonksiyonu X ve y verilerini test ve eğitim verileri olarak ayırır. test_size parametresi ile test verilerinin %20'si kullanılacağı belirtilir ve random_state parametresi ile de verilerin rastgele seçilmesi için belirlenir.

```
data_tahmin = ["I hate you","you are clever man","ugly bad girl"]
result=predict(cv,classfier_lr,data_tahmin)
print(result)
```

Fonksiyon, verilen metin verisi için "cv" (vektörleştirici) ve "classfier_lr" (lojistik regresyon sınıflandırma modeli) parametrelerini kullanır. Fonksiyon tarafından döndürülen veri çerçevesi, sonuçlarını içeren "result" değişkenine atanır ve sonuç olarak yazdırılır.

Çıktısı:

	tweet	target
0	I hate you	Offensive
1	you are clever man	Possitive
2	ugly bad girl	Hate

Data listesini Pandas veri çerçevesine dönüştürür. "pd.DataFrame()" fonksiyonu, verilen "data" listesini veri çerçevesine dönüştürür ve "columns" parametresi ile sütun başlıkları belirlenir. Daha sonra, veri çerçevesindeki sınıflandırma etiketleri (0, 1, 2) daha anlaşılır şekilde ("Hate","Offensive","Positive") ile değiştirilir. Son olarak, güncellenmiş veri çerçevesi fonksiyon tarafından döndürülür.

Logistik regresyon sınıflandırıcısı `classifier_lr` kullanarak hakaret içerikli, pozitif ve saldırgan olarak üç konu alanını çalıştırmaktadır. Aynı zamanda duygu analizi yapmaya çalışmaktadır. `Predict` fonksiyonu verilen metinlerin hakaret içerikli olup olmadığını ve duygusal tonunu tahmin etmek için kullanılır. Sonuçlar `result` değişkenine atanır ve `print(result)` ile ekrana yazdırılır.

```
y_pred_np = classifier_np.predict(X_test)
cm = confusion_matrix(y_test, y_pred_np)
print(cm)
```

Bu kod bloğu, "`classifier_np`" isimli bir sınıflandırma modelinin test verileri üzerindeki performansını ölçmek için kullanır. İlk olarak, model "`X_test`" verileri kullanılarak tahmin yapar ve tahmin sonuçları "`y_pred_np`" değişkenine atanır. Daha sonra, gerçek sınıf etiketleri (`y_test`) ve tahmin sonuçları arasındaki karışıklık matrisi oluşturulur ve ekrana yazdırılır. Karışıklık matrisi, modelin doğru ve yanlış tahminlerinin sayısını görmeyi sağlar.

```
y_pred_knn = classifier_knn.predict(X_test)
cm = confusion_matrix(y_test, y_pred_knn)
print(cm)
```

Bu kod bloğu, "`classifier_knn`" isimli bir KNN (K En Yakın Komşu) sınıflandırma modelinin test verileri üzerindeki performansını ölçmek için kullanır. İlk olarak, model "`X_test`" verileri kullanılarak tahmin yapar ve tahmin sonuçları "`y_pred_knn`" değişkenine atanır. Daha sonra, gerçek sınıf etiketleri (`y_test`) ve tahmin sonuçları arasındaki karışıklık matrisi oluşturulur ve ekrana yazdırılır.

```
y_pred_dt = classifier_dt.predict(X_test)
cm = confusion_matrix(y_test, y_pred_dt)
print(cm)
```

Bu kod bloğu, "`classifier_dt`" isimli bir Karar Ağacı sınıflandırma modelinin test verileri üzerindeki performansını ölçmek için kullanılır. İlk olarak, model "`X_test`" verileri kullanılarak tahmin yapar ve tahmin sonuçları "`y_pred_dt`" değişkenine atanır. Daha

sonra, gerçek sınıf etiketleri (y_test) ve tahmin sonuçları arasındaki karışıklık matrisi oluşturulur ve ekrana yazdırılır. Karışıklık matrisi, modelin doğru ve yanlış tahminlerinin sayısını görmeyi sağlar.

```
y_pred_lr = classifier_lr.predict(X_test)
cm = confusion_matrix(y_test, y_pred_lr)
print(cm)
```

Bu kod bloğu, "classifier_lr" isimli bir Lojistik regresyon sınıflandırma modelinin test verileri üzerindeki performansını ölçmek için kullanılır. İlk olarak, model "X_test" verileri kullanılarak tahmin yapar ve tahmin sonuçları "y_pred_lr" değişkenine atanır. Daha sonra, gerçek sınıf etiketleri (y_test) ve tahmin sonuçları arasındaki karışıklık matrisi oluşturulur ve ekrana yazdırılır. Karışıklık matrisi, modelin doğru ve yanlış tahminlerinin sayısını görmeyi sağlar.

```
y_pred_rf = classifier_rf.predict(X_test)
cm = confusion_matrix(y_test, y_pred_rf)
print(cm)
```

Bu kod bloğu, "classifier_rf" isimli bir random forest "rastgele orman" sınıflandırma modelinin test verileri üzerindeki performansını ölçmek için kullanılır. İlk olarak, model "X_test" verileri kullanılarak tahmin yapar ve tahmin sonuçları "y_pred_rf" değişkenine atanır. Daha sonra, gerçek sınıf etiketleri (y_test) ve tahmin sonuçları arasındaki karışıklık matrisi oluşturulur ve ekrana yazdırılır. Karışıklık matrisi, modelin doğru ve yanlış tahminlerinin sayısını görmeyi sağlar.

```
rf_score = accuracy_score(y_test, y_pred_rf)
lr_score = accuracy_score(y_test, y_pred_lr)
dt_score = accuracy_score(y_test, y_pred_dt)
np_score = accuracy_score(y_test, y_pred_np)
knn_score = accuracy_score(y_test, y_pred_knn)
```

scikit-learn kütüphanesindeki `accuracy_score` fonksiyonunu kullanarak farklı sınıflandırıcılar için doğruluk skorlarını hesaplamaktadır. Doğruluk skoru, sınıflandırıcının yapmış olduğu tahminlerin (`y_pred_rf`, `y_pred_lr`, `y_pred_dt`, `y_pred_np`, `y_pred_knn`) gerçek etiketlere (`y_test`) ne kadar yakın olduğunu ölçen bir metriktir. Sonuçlar, `rf_score`, `lr_score`, `dt_score`, `np_score`, `knn_score` değişkenlerine atanır.

```
print('Random Forest Accuracy: ', str(rf_score))
print('Logistic Regression Accuracy: ', str(lr_score))
print('Decision Tree Accuracy: ', str(dt_score))
print('Naive Bayes Accuracy: ', str(np_score))
print('KNN Accuracy: ', str(knn_score))
```

Hesaplanan doğruluk skorlarını ekrana yazdırmak için kullanılır. Her bir sınıflandırıcı için doğruluk skoru (`rf_score`, `lr_score`, `dt_score`, `np_score`, `knn_score`) ekrana yazdırılır ve hangi sınıflandırıcının ne kadar doğru olduğu görülebilir.

```
Random Forest Accuracy: 0.8694775065563849
Logistic Regression Accuracy: 0.890054468428485
Decision Tree Accuracy: 0.8432519669154731
Naive Bayes Accuracy: 0.5366148880371192
KNN Accuracy: 0.8297357272543877
```

SONUÇ

Bu çalışmanın amacı literatüre katkı sağlamak ve yapılan geliştirme sayesinde hakaret içerikli cümleler, saldırgan tavrı cümleler ve herhangi bir olumsuz tavrı olmayan cümleleri veri setine göre analiz edilip, hakaret içerikli cümlelerin ortaya çıkarılarak duygu analizi uygulanması sağlanmıştır. Duygu analizi ile insanların davranışlarla ilgili çıkarım yapma ihtimali artar. Bu yüzden yapılan çalışma gelişime açık olup farklı platformları da kapsayacak şekilde geliştirilebilir. Tez çalışması kapsamında beş farklı sınıflandırma algoritmaları kullanıldı. RF, LR, DT, NB, KNN algoritmaları kullanıldı. Etiketlenmiş veri seti üzerinden çıkan doğruluk (accuracy) oranlarına göre Logistic Regression modeli uygulandı. Kullanılan algoritmalar doğrultusunda veri setindeki belirtilen cümleleri tahmin ederek (predict) doğru sonuca ulaşması kapsamında LR modeli çalışmada uygun görüldüğü analiz sonucu ortaya çıkmıştır. Bu kapsamda Logistic Regression Accuracy değeri yaklaşık olarak 0.89 çıkarak, başarılı sonuç elde edildi. Üç parametre olarak belirlenen alanlara hate, offensive, none başlıklı yapılar kullanıldı. Cümleler yazıldığında kullanılan sınıflandırma algoritmasına göre tahmin (predict) etmekte sonrasında bu üç parametreden doğru olanı getirmektedir. Sosyal medya platformlarında aktif olarak sürekli veri akışı olduğu için bu platformlarda train, test olarak sınıflandırma algoritmalarına ayırmak daha kolaydır. Tez kapsamında adım adım işlem adımlarını doğru bir şekilde yapmaktır. İşlem adımlarını yaptıktan sonra önemli olan label data (etiketlenmiş veri) predict edip accuracy değeri yüksek olan sınıflandırma algoritmasını seçme durumu olacaktır. Piyasada kullanılan bir yapay zeka mekanizması olup; insanların duygusal durumunu ölçerek, bu parametreleri kullanıp duygu analizini ölçme hedeflenmektedir. Bu bağlamda insanların psikolojik açıdan duygularının analiz edilmesi mümkündür. Sosyal medya gibi platformlarda yapılan yorumlara istinaden duygu analizinin fazla olması sebebi ile bu platformlarda oldukça fazla ve etkili veri setinin olma olasılığı yüksektir. Ayrıca sosyal medya gibi platformlarda hakaret içerikli cümleler ve saldırgan tavrı cümleleri tespiti ile bu kişiler hakkında da yasal süreç başlatılabilir. Bu da ilgili kurumlarda bu yapının kullanılmasının ne kadar önemli olduğunu göstermektedir. Bu çalışmadan çıkarım olarak, hakaret içerikli cümlelerin tespiti ve duygu analizi gerekli algoritmalar kullanılması sonucunda ulaşılması mümkündür. Sosyal medya platformları, yapay zeka teknolojilerini kullanarak kullanıcılarının deneyimini geliştirmeye çalışmaktadır.

Burada amaç, kullanıcılara daha özgün ve kişiselleştirilmiş bir deneyim sunmak ve aynı zamanda sosyal medya platformlarına daha fazla kullanıcı çekmek ve onları mevcut kullanıcılar olarak tutmaktır. Yapay zeka teknolojileri, sosyal medya platformlarındaki çok sayıda kullanıcının verilerini analiz etmek ve yorumlamak için kullanılabilir. Bununla birlikte, yapay zeka teknolojilerinin sosyal medya platformlarında kullanımı bazı endişeleri de beraberinde getirme ihtimali olabilir. Örneğin, yapay zeka teknolojilerinin yanlış anlamaları sonucu insanların maruz kalabileceği hatalar, kullanıcılarda güvensizlik yaratabilir. Bu nedenle, sosyal medya platformları yapay zeka teknolojilerini kullanırken, bu teknolojilerin çok doğru bir şekilde çalıştığından emin olmak için çaba harcamalıdır. Yapılan çalışmalar neticesinde doğruluk oranları yüksek oldukça insanların bu teknolojilere güveni artarak yaygınlaşmasına olanak sağlar.

KAYNAKLAR

- Anderson, J., R. Michalski ve T. Mitchell. (1983). *Machine Learning: An Artificial Intelligence Approach*, M. Kaufmann
- Ataseven, B. (2013). Yapay Sinir Ağları İle Öngörü Modellemesi. *Öneri Dergisi*, 10(39), 101-115.
- Batrinca, B. ve Treleaven, P. C. (2015). Social Media Analytics: A Survey Of Techniques, Tools And Platforms. *Ai & Society*, 30, 89-116.
- Bello-Orgaz, G., J. Jung ve D. Camacho, (2016). Social Big Data: Recent Achievements And New Challenges. *Information Fusion*, 28, 45-59.
- Burcu, A. ve U.T.G. Şimşek. (2018). Sosyal Medya Analitiği İle Değer Yaratma: Duygu Analizi İle Geleceğe Yönelim. *Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 5(3), 797-811.
- Gürsoy, U. T. ve S. Bilgin. (2016). Banka Müsterilerinin Internet Bankacılığına İlişkin Yaklaşımlarının Veri Madenciliği Teknikleri İle İncelenmesi Kafkas University. *Faculty Of Economics And Administrative Science*, 7(14), 421-442.
- Güvenç, E., M. Sakal, G. Çetin ve O. Özkaraca. (2022). Öğrencilerin Dersteki Niteliklerinin Makine Öğrenmesi Teknikleri Kullanılarak Sınıflandırılması. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 10(3), 1359-1371.
- Harman, S. (2020). *Makine Öğrenmesi | Clustering (Kümeleme) Teknikleri*, 15 Ocak 2023 tarihinde <https://samed-harman.medium.com/makine-%C3%B6%C4%9Frenmesi-clustering-k%C3%BCmeleme-teknikleri-bd1b59a0a177> adresinden alındı.
- James, G., D. Witten, T. Hastie ve R. Tibshirani. (2013). An Introduction To Statistical Learning. *Springer*, 316-321.

Kaplan, A. M. ve M. Haenlein. (2010). Users Of The World, Unite! The Challenges And Opportunities Of Social Media. *Business Horizons*, 53(1), 59-68.

Kumaş, E. (2021). Türkçe Twitter Verilerinden Duygu Analizi Yapılırken Sınıflandırıcıların Karşılaştırılması. *Eskişehir: Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 2(2), 1-5.

Maltarollo, V., G.K.M. Honório ve A.B.F. Da Silva. (2013). Applications of Artificial Neural Networks in Chemical Problems. *Artificial Neural Networks-Architectures And Applications*, 203-223.

Michie, D., D. J. Spiegelhalter, C. C. Taylor ve J. Campbell (1994). Taylor, Machine Learning Neural and Statistical Classification, *New York: Ellis Horwood Limited*.

Nalçakan, Y., Ş. S. Bayramoğlu ve S. Tuna. (2015). Sosyal Medya Verileri Üzerinde Yapay Öğrenme ile Duygu Analizi Çalışması. *Technical Report*.

Nemschoff, M. (2013). Social Media Marketing: How Big Data Is Changing Everything. *Cms Wire*, 16.

Öztemel, E. (2006). *Yapay Sinir Ağları. 2. Baskı*. İstanbul: Papatya Yayıncılık.

Öztürk, K., M.E. Şahin. (2018). Yapay Sinir Ağları Ve Yapay Zekâ'ya Genel Bir Bakış. *Takvim-i Vekayi*, 6(2), 25-36.

Piryonessi S. M. ve T. E. El-Diraby. (2020). Role Of Data Analytics In Infrastructure Asset Management: Overcoming Data Size And Quality Problems. *Journal Of Transportation Engineering, Part B: Pavements*, 146, 1-17.

Samoshtn, A. (2020). Veri Seti, 10 Aralık 2022 tarihinde <https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset> adresinden alındı.

Savaş, S., N. Topaloğlu, ve M. Yılmaz. (2012). Veri madenciliği ve Türkiye’deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21), 1-23.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization, *ACM Computing Surveys (CSUR)*, 34, 1-47.

Shen, L., H.Chen, Z. Yu, W.Kang, B. Zhang, H. Li ve D. Liu. (2016). Evolving Support Vector Machines Using Fruit Fly Optimization For Medical Data Classification. *Knowledge-Based Systems*, 96, 61-75.

